PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/60

A1

(11) International Publication Number: WO 00/26839

(43) International Publication Date: 11 May 2000 (11.05.00)

(21) International Application Number: PCT/US99/26083

(22) International Filing Date: 3 November 1999 (03.11.99)

(30) Priority Data:

60/107,063	4 November 1998 (04.11.98)	US
PCT/US98/27664	28 December 1998 (28.12.98)	US
09/380,219	27 August 1999 (27.08.99)	US

(71) Applicant *(for all designated States except US)*: INFODREAM CORPORATION [-/US]; 2340A Walsh Avenue, Santa Clara, CA 95051 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: ANDLEIGH, Prabhat, K. [-/US]; 10701 Castine Avenue, Cupertino, CA 95014 (US). PAPPU, Nagaraju [-/US]; Apartment 14 H, 20800 Homestead Road, Cupertino, CA 95014 (US). KALINDINDI, Vasudeva, V. [-/US]; Apartment 95, 3655 Pruneridge Avenue, Cupertino, CA 95051 (US).

(74) Agents: RADLO, Edward, J. et al.; Fenwick & West LLP, Two Palo Alto Square, Palo Alto, CA 94306 (US).

(81) Designated States: CA, GB, IN, US.

Published

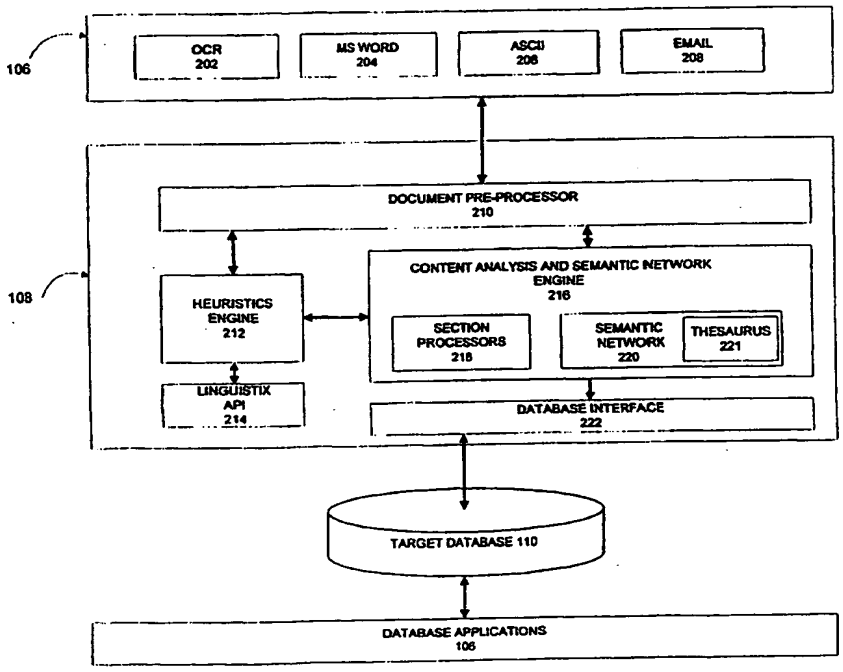
With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: ADVANCED MODEL FOR AUTOMATIC EXTRACTION OF SKILL AND KNOWLEDGE INFORMATION FROM AN ELECTRONIC DOCUMENT

(57) Abstract

An apparatus, method, and computer readable medium for analyzing and extracting skill and knowledge information from an electronic document (104) and for storing the extracted skill and knowledge information into predefined fields or tables in a target database (110) comprises a content analysis and semantic network engine (216) for analyzing and extracting skill and knowledge information from the electronic document (104). A skill and knowledge information extractor (702) is coupled to the content analysis and semantic network engine (216), for determining a skill level for the skill information extracted from the electronic document (104). In a preferred embodiment, the skill and knowledge section processor (702) uses a non-monotonic reasoning principle to determine a skill level for skill information extracted from the electronic document (104). The content analysis and semantic network engine (216) further comprises a thesaurus (221) for linking together terms (402) and skill information (404), and for defining relationships between and among the terms (402) and skill information (404), and a semantic network (220) coupled to the thesaurus (221), for organizing the terms (402) and skill information (404) in the thesaurus (221), along with knowledge information (502) and categories (504), in a hierarchical structure.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

ADVANCED MODEL FOR AUTOMATIC EXTRACTION OF SKILL AND KNOWLEDGE INFORMATION FROM AN ELECTRONIC DOCUMENT

RELATED APPLICATION

The subject matter of this application is a continuing application of and claims priority from U.S. patent application Serial No. 09/380,219, filed August 27, 1999 descending in priority from PCT application PCT/US98/27664, filed on December 28, 1998, and entitled "Xtraction Server" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi. Said two earlier applications are commonly assigned with the instant application.

The subject matter of this application is also related to and claims priority from U.S. Provisional Application Serial No. 60/107,063, filed Novmeber 4, 1998, and entitled "Advanced Model for Automatic Extraction of Content, Skills, and Knowledge from Resumes" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi, which application is commonly assigned with the instant application, and is incorporated herein by reference in its entirety.

TECHNICAL FIELD

This invention relates to the field of computer analysis of electronic documents. More specifically, it relates to the field of information retrieval to convert and store information in documents written in a natural language into a predefined structure which can be retrieved and manipulated by computer program applications.

BACKGROUND OF THE INVENTION

Information to be sorted and stored in a computer database may reside in numerous electronic documents. For example, information about people and their

specific talents and skills may reside in electronic documents, such as resumes, performance appraisals, design documents, publications, books, patent documents, and email messages. When an individual is trying to organize and sort out specific information from such electronic documents, the individual usually has to open each document separately and manually analyze, retrieve, and store the relevant data in the particular database. For example, a project manager who would like to find the best employee for a specific job may have a specific job description. When searching for an employee whose skills, knowledge and talent are best suited for the specific job description, the project manager must sift through several documents which contain the necessary information. Such a process is time consuming and inefficient, because the project manager may have to read the documents several times and may have to review and type the information into a computer database in order to organize the various pieces of information into a coherent summary.

A computerized system which can analyze and extract pertinent information from different electronic documents would provide a more efficient solution to this problem. However, such text documents are often written in unstructured natural language text for other people to understand. Thus, computer programs such as database applications cannot efficiently process documents written in natural language texts. Rather, computer programs can process only information which has been stored in a highly structured fashion in order to retrieve and manipulate that information. Additionally, these documents may be prepared in a variety of different file formats,

such as Microsoft Word 97, Rich Text Format, PDF, WordPerfect, ASCII files, and HTML, and may be stored in different areas within a computer.

There are a variety of information retrieval programs such as Internet search engines that can retrieve documents that match a set of keywords. Their scope is very limited in the context of the above mentioned problem, because they cannot understand the text, and certainly they cannot make any connection between the document and the person who is related to that document. Another problem is that the 'information of interest' will vary significantly from one organization to another. For example, a health care organization will be interested in the skills and talents related to the medical field, but the skills related to computers may not be of significant interest, whereas a software development organization will be interested in the computer and software related skills, but may not be interested in medical or first-aid related skills. The keyword based search engines cannot address this problem of retrieving only the 'information of interest'. As a result, there is a vast amount of information about people which cannot be easily processed by computer programs.

For example, in today's large corporations and government organizations, it is not uncommon to receive hundreds of thousands of resumes of potential candidates in a very short time. Recruiting the right candidates from such a vast pool of applicants is a very complicated problem. It is crucial for organizations to find the people with the right knowledge and skill set. In essence, managers have to deal with a vast number of resumes, try to understand the content within the resumes, and short-list candidates who have the right skills and knowledge. For example, if an organization wants to

recruit a middle level manager with 5 to 8 years of experience to lead a development project, the organization will need to sort through thousands of resumes and determine from each one whether that particular candidate has the requisite knowledge and skill level. It is not possible to find the best resumes using a standard full text search engine because such search programs search for a particular input string and retrieve only resumes which contain that particular input string. Such an approach is not that useful, because a particular skill may be written using many different terms (e.g. Microsoft Word, MS Word, Word 97, etc....) even though the terms all refer to the same or similar ideas. Moreover, in addition to not being able to correctly identify a candidate's skills, a typical search program cannot identify the type of experience with that skill, the duration of that experience, or the overall knowledge gained by the candidate in a specific skill group. Additionally, it is also very desirable to have a system for determining not only the knowledge and skills of a candidate but also the proficiency level of a candidate in a particular skill.

Therefore, what is needed is a system for analyzing and extracting information from an electronic document and for storing the extracted information in a database. Additionally, what is needed is a system for analyzing and extracting skill and knowledge information from an electronic document and for determining a skill level for skill information and for mapping such skill level information to a qualitative scale.

DISCLOSURE OF INVENTION

The present invention is an apparatus, method, and computer-readable medium for analyzing and extracting skill and knowledge information from an electronic

document (104) and for storing the extracted skill and knowledge information into predefined fields or tables in a target database (110). The system for analyzing and extracting skill and knowledge information from an electronic document (104) comprises a content analysis and semantic network engine (216) for analyzing and extracting skill and knowledge information from the electronic document (104), and a skill and knowledge information extractor (702) coupled to the content analysis and semantic network engine (216), for determining a skill level for the skill information extracted from the electronic document (104). In a preferred embodiment, the skill and knowledge section processor (702) uses a non-monotonic reasoning principle to determine a skill level for skill information extracted from the electronic document (104). The content analysis and semantic network engine (216) further comprises a thesaurus (221) for linking together terms (402) and skill information (404) and for defining relationships between and among the terms (402) and skill information (404), and a semantic network (220) coupled to the thesaurus (221), for organizing the terms (402) and skill information (404) in the thesaurus (221), knowledge information (502), and categories (504) in a hierarchical structure.

A method for extracting skill and knowledge information from an electronic document (104) comprises the steps of: identifying skill and knowledge information in the electronic document (802); determining a skill level for skill information from the electronic document (804); and mapping the skill level to a qualitative scale (806). The method further comprises the step of storing the skill information and qualitative skill level scale mapping in the target database (808).

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram of a preferred embodiment of a system 100 in accordance with the present invention.

Figure 2 is a block diagram of a preferred embodiment of an extraction server 108 in accordance with the present invention.

Figure 3 is a flow chart of a preferred embodiment of the steps performed by the document pre-processor 210.

Figure 4 is a block diagram of a preferred embodiment of a thesaurus. 221

Figure 5 is a block diagram of a preferred embodiment of a semantic network 220.

Figure 6 is a flow chart of a preferred embodiment of the steps performed by the extraction server 108.

Figure 7 is a block diagram of a preferred embodiment of a system 700 in accordance with the present invention.

Figure 8 is a flow chart of a preferred embodiment of the steps performed by the skill and knowledge information extractor 702.

Figure 9 is a screen shot of a user interface of a preferred embodiment of a target database 110 display for skill information.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring now to Figure 1, a system 100 upon which a preferred embodiment of the present invention operates is shown. A host computer 102, using the method and system described herein, operates upon an electronic document 104, derived from

a text document which contains unstructured text. As used herein "unstructured text" refers to any document which has been written in a natural language such as English. Examples of documents containing unstructured text include, but are not limited to, a resume, performance appraisals, design documents, publications, books, patent documents, and email messages. In a preferred embodiment, the host computer 102 is a conventional computer having a keyboard and mouse for input (not shown), and a conventional memory 106 associated with host computer 102 for storing the electronic document 104. The electronic document 104 may be prepared in any electronic file format, such as Microsoft Word 97, Rich Text Format, PDF, WordPerfect, ASCII files, and HTML.

The electronic document 104 is processed by host computer 102 using the present invention. Specifically, host computer 102 uses extraction server 108 to analyze, retrieve and store words and word groups from the electronic document 104 into a predefined structure in target database 110. As used herein, the terms "words" and "word groups" are used to mean any text that may be derived from document 104 including, but not limited to, individual words or numbers, phrases, whole sentences, and blocks of text. The extraction server 108 identifies the document type of the document 104 and determines which words and word groups are to be extracted from the document 104. The structure and operation of the extraction server 108 is described in more detail below with reference to Figures 2 through 6.

The target database 110 comprises predefined tables with predefined columns for storing the word and word groups extracted from the electronic document 104. In a

preferred embodiment, a predefined table and predefined columns correspond to a particular document type. For example, if document 104 is a resume, then a predefined table for a document type called "resume" may have predefined columns such as "name and address", "education", and "skills and experience". As another example, if document 104 is a patent document, then a predefined table for a document type called "patent document" may have predefined columns such as "inventors", "company", "patent number", and "field of search". The predefined tables and columns in target database 110 are organized ahead of time, and one skilled in the art will realize that the present invention is not limited to a particular document type or a predefined table, but that many different compilations of predefined tables and columns may be stored in target database 110 within the scope of this invention. The words and word groups stored in the target database 110 can be stored in electronic form on any type of computer data storage device or they may be printed out in a hard-copy printed format.

The process of extraction performed by the extraction server 108 preferably uses a non-monotonic reasoning principle. As used herein, a "non-monotonic reasoning principle" refers to a process whereby at every stage during extraction, the extraction server 108 assumes a reasonable default value. That default value is modified as further information becomes available. For example, a string '1987' is first assumed to be a number, and if further information to qualify the string to be a date is available (for example in this case, that the string is preceded by another string 'Jan'), then the assumption is changed. If again further information becomes available to negate the previous assumption, the assumption is changed again.

Thus, the present invention advantageously allows a user to extract skill and knowledge information from an electronic document directly into a database. More specifically, the present invention analyzes an electronic copy of a text document and extracts words and word groups relating to skill and knowledge information into a target database comprising predefined tables and columns associated with a particular document type. Moreover, the present invention operates upon electronic documents in any electronic file format. The extracted skill and knowledge information stored in the target database can then be retrieved and manipulated by other computer program applications.

Referring now to Figure 2, a block diagram of a preferred embodiment of the extraction server 108 is shown. The electronic document 104 may be any electronic file stored in memory 106 which is accessible by the extraction server 108. For example, the electronic document 104 may be an electronic form of a hard copy of a document converted using a conventional optical scanner and Optical Character Recognition (OCR) software 202, a Microsoft Word file 204, an ASCII text file 206 or an email attachment 208. The database applications which manipulate the extracted information in target database 110 are also preferably stored in memory 106. In a preferred embodiment, the extraction server 108 comprises a document preprocessor 210 coupled to the memory 106 where the electronic document 104 is stored, a heuristics engine 212 coupled to the document pre-processor 210, a morphological analysis engine 214 coupled to the heuristics engine 212, a content analysis and semantic network engine 216 coupled to the document preprocessor 210, and a

database interface 222 coupled to the content analysis and semantic network engine 216 and to the target database 110. The content analysis and semantic network engine 216 preferably comprises section processors 218 and a semantic network 220.

The document pre-processor 210 retrieves the electronic document 104 from memory 106 and performs the initial analysis of the electronic document 104. Referring now to Figure 3, a flowchart of the steps of a preferred operation of the document pre-processor 210 is shown. The document pre-processor 210 performs the initial analysis and extraction of the electronic document 104 by first converting (302) the electronic document 104 from its native file format into ASCII text. More specifically, the document pre-processor 210 identifies the file format of the electronic document 104 and extracts the ASCII text out the document 104. For example, if the electronic document 104 is a Microsoft Word file, then the document pre-processor 210 identifies the file by the Microsoft Word signature and uses the Microsoft Object Linking and Embedding Software Development Kit (Microsoft OLE 2.0 SDK) to extract text from the Microsoft Word File.

Next, the document pre-processor 210 filters out (304) any unnecessary and unwanted information such as, but not limited to, email headers, OCR headers, blank pages, and unwanted characters. Preferably, any information that is not part of the original document is treated as unnecessary information. For example, email headers, non-ASCII characters at the beginning or at the end of the file, extra blank lines and blank spaces are removed from the text. Additionally, if the text contains vertical tables, these tables are preferably converted into horizontal tables. If the text contains

multiple columns, it is preferably converted into single column. The document pre-processor 210 then stores (306) formatting information for the document 104 such as, but not limited to, the fonts used, font sizes, section tittles, and subsections.

The document pre-processor 210 then performs paragraph identification heuristics (308) on the electronic document 104. During this step, the beginning and end of each paragraph is identified, and the paragraph characteristics are gathered. As used herein, the phrase "paragraph characteristics" refers to the statistical properties of the paragraph. Paragraph characteristics include, but are not limited to, the number of words in the paragraph, the number of lines in the paragraph, the average number of words per line, whether any line has a bullet as the starting character, and whether there are any underlined sentences in the paragraph.

Finally, the document pre-processor 210 performs paragraph grouping heuristics (310) on the electronic document 104. Once the paragraphs have been identified, the document pre-processor 210 groups the paragraphs into sections. During this step, the paragraphs are grouped into sections based on the paragraph characteristics as well as using any section tittles that precede the paragraphs. Starting at the beginning of the electronic document 104, the first heading or section title is identified, and the following paragraphs until the next section title are grouped into one section. If no section titles are found, then using the paragraph characteristics, all the similar paragraphs are grouped into sections. Additionally, paragraphs that have same or similar characteristics are grouped together into sections.

The heuristic engine 212 applies a set of heuristics, that is a set of rules, to the electronic document 104 for analyzing information in the electronic document 104. The set of heuristics which are applied to the electronic document 104 are associated with a particular document type. For example, if the document type is a "resume", then the set of heuristics associated with the document type "resume" is applied to the electronic document 104. Heuristics are described below in more detail in commonly assigned U.S. patent application Serial No. 09/380,219 entitled "Extraction Server" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi, which is incorporated herein by reference in its entirety.

The morphological analysis engine 214 is used for target language analysis and is preferably the LinguistiX 2.0 application programming interface (API) from InXight Corporation in Palo Alto, CA. The LinguistiX 2.0 API is a language neutral programming interface. In other words, the LinguistiX API can analyze documents in any language such as English, French or German. Because the heuristics engine 212 and the LinguistiX API are external to and separate from the document pre-processor 210 and the content analysis and semantic network engine 216, the present invention can extract information from documents in the English, French or German language, and any other languages which will be supported by the LinguistiX API in future.

Preferably, the Heuristics Engine 212 uses the following features provided by the LinguistiX API: tokenization, lexical analysis, tagging, and noun-phrase extraction. Before text from the electronic document 104 can be analyzed in terms of its linguistic roots and function, it must first be segmented into words, punctuation and idiomatic

phrases. LinguistiX tokenization includes the ability to recognize multi-word constructs such as HTML tags. The lexical analysis feature identifies the grammatical features of a word in addition to its root forms. The tagging feature identifies the grammatical category of words by their context. The noun-phrase extraction identifies multi-word phrases in documents. LinguistiX phrase extraction technology enables software to work with these larger concepts to provide improved information analysis and retrieval. For example, 'Windows Programming' will be identified as one phrase, instead of two distinct words Windows and Programming. This feature is used by the semantic network 220 to identify the multi-word noun phrases.

These features of the LinguistiX API are used to implement the heuristics. For example, by using the tagging feature, the extraction server 108 may discover that a particular word is a proper noun. Whether that word is the name of the person or the name of a company will depend on where the word occurred in a document. For example, if the word occurs in a contact information section of a document, then it may be the name of the person, or name of the street, city and so on. If the word occurs in an experience section of a document, and if it is followed by the name of a city and state, it may be a company name.

The database interface 222 is a set of APIs that provide a mechanism for retrieving and storing information to and from the target database 110. This is done in such a way that the underlying implementation of the target database 110 is hidden from the application using the database interface. Thus, the extraction server 108 can work with any industry standard relational database software such as Oracle or

Microsoft SQL Server without having to change the software or its implementation. Additionally, the database interface 222 provides the following mechanisms: a method to connect to the target database, a method to maintain the connection to the database, a transaction model to maintain the consistency of the database, and various methods to retrieve, query, update, insert and delete information from the target database 110.

The content analyzer and semantic network engine 216 analyzes the content of the electronic document 104, extracts words and word groups from the document 104, and stores the extracted information in the appropriate tables in the target database 110. In a preferred embodiment, the content analyzer and semantic network engine 216 comprises section processors 218 which extract information from a particular section of interest, and a semantic network 220. The semantic network 220 uses a thesaurus 221 and a phrase extraction process to identify the meta-concepts and categories in the electronic document 104 and extracts related words and word groups into the target database 110. In a preferred embodiment, the present invention may be implemented to run on a Windows NT Server and Oracle Database.

Referring now to Figure 4, a block diagram of a preferred embodiment of a thesaurus 221 is shown. The thesaurus 221 is a vocabulary database for the extraction server 108 and is organized by skills. The thesaurus 221 groups all related terms 402 in a language under a language independent concept 404. As used herein, a "term" 402 refers to all the individual words or word groups that belong to a particular language along with their alternatives. As used herein, a "concept" or "skill" 404 comprises a set of terms 402 that are language specific and alternatives to one another. However,

the skill 404 itself is language independent. Skills 404 establish synonymous relationships among all terms 402 in the thesaurus 221 that have the same meaning. In other words, skills 404 connect all the different names for the same skill 404 that are known to the thesaurus 221 and specify certain characteristics for each name. Preferably, each skill 404 has a unique skill identifier (ConceptID). The Concept ID by itself has no intrinsic meaning. Each term 402 in each language in the thesaurus 221 has a unique term identifier. The same term 402 in different languages, for example, in English and Spanish, will have a different term identifier for each language.

To illustrate the relation between terms 402 and skills 404 consider an example in which term1 402A may consist of 'MS VC++', term2 402B may consist of 'Microsoft Visual C++' and term3 402C may consist of 'MS Visual C++'. All these terms 402 are linked to the skill 404 'Visual C++'. In other words, if the electronic document 104 uses any of the words or word groups 'MS VC++', 'Microsoft Visual C++' or 'MS Visual C++', the thesaurus 221 allows the extraction server 108 to recognize the words or word groups as being linked to the skill1 404A 'Visual C++'. In another example, term4, term5 and term6 are respectively 'JDK 1.1', 'Symantec Café', and 'JDBC', and all these terms 402 are linked to the skill2 404B called 'Java'. Thus, if the electronic document 104 uses any of the words or word groups 'JDK 1.1', 'Symantec Café', and 'JDBC', the thesaurus 221 allows the extraction server 108 to recognize the word or word group as being linked to the skill2 404B 'Java'.

The thesaurus 221 may also comprise other information such as the attributes of a skill 404 or attributes of a term 402. Attributes provide additional information that helps to define the meaning of a skill 404 and explain how it may be used in a document. In other words, the different senses of a particular word or word groups are captured using the attributes.

In addition to the relationship between a skill 404 and a set of terms 402, the thesaurus 221 also comprises relationships among skills 404. Preferably, these relationships are non-subsumption relationships. As used herein, the term "non-subsumption" refers to relationships that include related skills, co-occurring skills and/or associated expressions. In other words, non-subsumption refers to relationships that are not based on subsumption. For example, C++ and Java are related, but neither subsumes the other. All these relationships among skills 404 indicate that the skills 404 linked together are not exactly similar but are associated with each other in different ways. One skilled in the art will realize that the terms and skills of the thesaurus 221 are not limited to the examples given herein but may contain any number of terms and skills which have been predefined and stored in the thesaurus 221 prior to the processing of the electronic document 104. Thus, the thesaurus advantageously allows the present invention to link together terms and skills used in specific industries, disciplines, and technologies for which the thesaurus is being used, and preserves the meanings and hierarchical connections between those terms and skills. Additionally, the thesaurus facilitates the access to concept relationships and to term and skill attributes irrespective of the term used as a point of entry.

Referring now to Figure 5, a block diagram of a preferred embodiment of a semantic network 220 is shown. The semantic network 220 provides a way of arranging all the skills 404 at the lowest level and then builds a taxonomy or network of higher level knowledge-concepts and categories. The semantic network 220 comprises skills 404 at the lowest level, "knowledge" or knowledge-concepts 502 at a second level, and categories 504 at the highest level. The semantic network 220 together with the thesaurus 221 provides a four level hierarchy of terms 402, skills 404, knowledge-concepts 502 and categories 504.

A category 504 is the highest level in the semantic network 222. Broad categories 504 may be created according to a specific industry which fully subsume other knowledge-concepts 502 and skills 404. The semantic network 220 categorizes all knowledge-concepts 502 into categories 504. Knowledge-concepts 502 comprises the next level in the semantic network 220 hierarchy. Each knowledge-concept 502 is a collection of skills 404 that add to the body of knowledge. The semantic network 220 categorizes all skills 404 into knowledge-concepts 502. As described earlier with reference to Figure 4, skills 404 are generic and language independent from all related terms 402. The semantic network 220 categorizes all terms 402 into skills 404. As described earlier with reference to Figure 4, terms 402 comprise language dependent strings that are found in the electronic document 104. Terms 402 comprise the lowest level in the semantic network 220 hierarchy.

The entire semantic network 220, separate from the thesaurus 221, comprises language independent knowledge that is arranged as a taxonomy. Preferably, the

relationships between skills 404 and knowledge-concepts 502 as well as the relationships between knowledge-concepts 502 and categories 504 are many to many. In other words, a single knowledge-concept 502 can comprise several skills 404 and a single skill 404 can be linked to several knowledge-concepts 502. Similarly, several knowledge-concepts 502 may comprises a category 504 and several categories may have links to a single knowledge-concept 502.

To illustrate the terms 402, skills 404, knowledge-concepts 502, and categories 504 of a semantic network 220, the two concepts discussed earlier with reference to Figure 4, namely 'Visual C++' and 'Java', will be used. Both these skills 404 may be grouped under a knowledge-concept 502 'Object Oriented programming languages'. Additionally, the skill 404 'Visual C++' may also belong to the knowledge-concept 502 'Visual Programming Environment'. The knowledge-concept 502 "Visual Programming Environment" may also be linked to other skills 404 such as 'Visual Basic'.

The semantic network 220 uses subsumption as the basis for the hierarchical organization of skills 404, knowledge-concepts 502, and categories 504. In other words, the relationship between skills 404 and knowledge-concepts 502 and knowledge-concepts 502 and categories 504 in the semantic network 220 are based on conceptual subsumption, where a more general object 'subsumes' a more specific object. The concept of subsumption is more general than the concept of synonymy. An object is subsumed by another object if the subsuming object is much more general than any other subsumed objects and effectively summarizes the subsumed objects.

Truly synonymous objects mutually subsume each other. If only synonymous based relationships are allowed, then the granularity between the objects cannot be captured effectively as there are not many truly synonymous objects. The difference between the shades of meaning will not allow correct retrieval in a synonym-based network. The subsumption-based network removes these drawbacks and aids in retrieving related concepts more accurately, since a subsumption is more general compared to a synonym. For example, the object 'JDBC' is subsumed by a more general object called 'Java Programming Language' (a knowledge-concept 502), which is further subsumed by an even more generic object 'Software Engineering' (a category 504).

An object may also be subsumed by more than one higher level object. For example, the skill 404 'JDBC' may be subsumed by at least two knowledge-concepts 502 such as 'Java Programming Language' and 'Database Connectivity Library'. Each of these knowledge-concepts 502 may in turn be subsumed by several categories 504. Hence, the conceptual subsumption also allows many-to-many relationships between skills 404 and knowledge-concepts 502 and between knowledge-concepts 502 and categories 504.

Referring now to Figure 6, a flowchart of the steps of a preferred embodiment of a method performed by the content analysis and semantic network engine 216 is shown. First, identification heuristics are performed (602) on the electronic document 104 to identify the beginning and end of the known sections of interest. The sections of interest are configured by the user when the extraction server 108 is first installed. The sections are then analyzed (604) and information is extracted from the sections. The

extracted information is stored (606) in a predefined structure in the target database 110. Using the semantic network 220, words and word groups are analyzed (608) and the relationships between the different words and word groups are determined and stored in the target database 110. Thus, the present invention advantageously extracts meaningful information from electronic documents, and stores them in a predefined structure in a target database. The extracted information stored in the target database can then be retrieved and manipulated by computer program applications accessing the database. Moreover, the present invention provides a powerful semantic network and thesaurus for defining terms, concepts, meta-concepts, and categories and the relationship between and among such terms, concepts, meta-concepts, and categories. Thus, the semantic network can stored information relating to any field, industry or technology, and allows the extraction server 108 to process various types of documents pertaining to such fields, industries or technologies.

The section processors 218 extract information from sections of interest in an electronic document 104. The particular sections of interest from which information is extracted is determined by the document type. The content analysis and semantic network engine 216 comprises a section processor 218 for extracting words or word groups from each section of interest in an electronic document.

Section processors 218 are configured to operate on a specific document type and may contain one or several section processors 218. For example, resumes typically contain several sections such as a cover letter, contact information, an objective section, an experience section, an education section, a patents section, a

publications section, an awards and honors received section, and a courses attended section. In a preferred embodiment, section processors 218 for a resume document type may comprise a cover letter section processor for extracting information from a cover letter, a contact information section processor for extracting contact information for a candidate, a skills and experience section processor for extracting the skills and experience of a candidate, an education section processor for extracting educational information from a candidate, an awards and honors section processor for extracting any awards and honors received by a candidate, a patents section processor for extracting information about patents obtained by a candidate, and a publications section processor for extracting any articles or documents published by a candidate. Each section processor 218 analyzes a particular section in the electronic document 104 and extracts specific words and word groups from that section into a specific record in the target database 110. Additionally, as described in more detail in commonly assigned U. S. Patent Application Serial No. 09/380,219 entitled "Xtraction Server" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi, each section processor 218 applies a set of heuristics to the particular section of interest in order to analyze and extract the desired information.

Referring now to Figure 7, there is shown a preferred embodiment of the present invention comprising a skills and knowledge information extractor 702. The skills and knowledge information extractor 702 allows the system to automatically extract from a document, such as a resume, the skills of a candidate, the candidate's knowledge in a particular area, and to determine the proficiency level of the candidate

in any given skill. Thus, the skills and knowledge information extractor 702 allows a user to automatically determine a "career profile" of a candidate from his or her resume. As used herein, a "career profile" refers to any qualitative and quantitative information about a candidate's work history, experience, and proficiency. For example, such information includes, but is not limited to, how long a candidate worked in a particular profession, when, where, and at what depth did the candidate gain experience in a particular skill, what is the candidate's overall knowledge level in a particular area, how much management experience a candidate has, etc.

As used herein, "terms" refers to the actual word or words which are found in a resume, "skill" or "skill information" refers to the skills 404 in the thesaurus 221 and semantic network 220 which relate to those terms, and "knowledge" or "knowledge information" refers to the knowledge-concepts 502 relating to the skills. For example, in a resume, a candidate may have used the terms "Microsoft Visual C++" or "MS VC++". The present invention would identify these terms as belonging to the skill "C++", which in turn is related to the knowledge "object oriented programming" which in turn may be related to the category "Software." Thus, although the only terms actually used in and extracted from the resume were "Microsoft Visual C++" and "MS VC++", the present invention is able to determine that the candidate has "skill" in C++ and has "knowledge" of object oriented programming even though the words C++ and object oriented programming were never used in the document.

The skill and knowledge information extractor 702 uses a non-monotonic reasoning principle to determining a candidate's skill level. As described above, non-

monotonic reasoning refers to the use of default assumptions which are made about the state of unknown factors. These default assumptions may be changed as new information or evidence becomes available. Additionally, default assumptions may be changed due to the absence of certain information or evidence. The operation of the non-monotonic reasoning approach used by the skill and knowledge for information extractor 702 is best illustrated using an example.

During operation, the present invention finds a skill, X, in a candidate's resume, R. In the absence of any other knowledge, the skill and knowledge information extractor 702 assumes that the skill level of the candidate for the skill X is average. As the skill and knowledge information extractor 702 obtains additional information from the resume R about skill X, the assumption of the skill level for skill X is refined. Additional knowledge that may be used to refine the skill level includes, but is not limited to, the section in which the skill X is found. For example, if the skill X is found in the Objective Section of a resume R, a positive numerical value, or objective weightage factor W(O), will be added to the skill level. Additionally, a positive weight for each project in which the skill X is used, represented here by $W(P_i)$, may be added to the skill level. Preferably, this weightage value is computed for all projects in resume R. The number of associated skills that are also used, W(K), may also be added to the skill level. As used herein, associated skills are the skills related to the main skill; knowing a main skill implies that a person also knows all associated skills. For example, if one is an expert in the skill "database programming" or "database administration," this person must be knowledgeable in the associated skill

"SQL." Associated skills can be determined using the semantic network 220 and the thesaurus 221. For a given skill x, all its associated skills ($X_1 \ldots X_n$) are linked with x through the semantic network 220 and thesaurus 221. For example, a thesaurus 221 entry for the "skill database administration" would contain links to the "skills database server administration," "database user management," and/or "SQL." Also, the number of years of experience for the skill X, W(Y) may also be added to the skill level. Moreover, the number of years since the skill X was used may represent a negative factor, W(LU), which is subtracted from the skill level. Thus, in a preferred embodiment, a summation of the weights described above gives a specific skill level for the skill X. A mathematical representation for determining the skill level of a particular skill is as follows:

$$\text{SkillLevel}(X') = \text{SkillLevel}(X) + W(O) + \sum W(P_i) + W(K) + W(Y) - W(LU)$$

The weightage functions are computed using the total number of skill levels that are defined, and the distance from the current skill level to the next skill level. One skilled in the art will realize that the weightage factors used to adjust the skill level are not limited to those listed in the above example but can comprise any number of factors to be determined by the system creator.

The computation of the skill level of a particular skill for a candidate can also be demonstrated using an example. Initially, the skill and knowledge information extractor 702 assumes that a person has an average skill level for a particular skill such

as C++. If the candidate's resume states that the candidate took a course in C++, that fact would add a positive weightage factor to the skill level, thus adjusting the average skill level to a higher value. If the candidate's resume also states that the candidate has two years of work experience in C++, that fact would add another positive weightage factor to the skill level and adjust the average skill level to another higher value. The values by which the average skill level is adjusted for the C++ course and the two years of work experience are not necessarily the same but may reflect the value attributed by the system creator. Each mention of C++ in the resume would this be used to adjust the skill level either up or down. Additionally, the user of terms in the resume which are related in the semantic network 220 and thesaurus 221 to the concept or skill C++ could also be used to adjust the skill level of the candidate. After all the relevant terms in the candidate's resume have been extracted and evaluated, the skill and knowledge information extractor 702 determines a single value for the skill level for the candidate for the particular skill.

After a final skill value for a particular skill has been determined, the skill and knowledge information extractor 702 then maps the skill value to a scale for qualitatively illustrating the proficiency of the candidate in that particular skill. For example, if a final skill value for a particular candidate has been determined to be the number 6.8, that number may map to a rating of "good" on a scale of 1 to 10, with 1 being poor and 10 being excellent. Thus, the present invention allows a user to determine the proficiency of a candidate's skill level for a particular skill and to ascribe a qualitative value to that proficiency level. One skilled in the art will realize that the

qualitative scales used to describe a particular skill value may be any type of scale with a range of numerical values and/or adjective descriptors. For example, a qualitative scale may map the final skill value to a scale comprising numbers such as 1 to 5 or 1 to 10. A scale may map the final skill value to a scale comprising numbers and adjectives such as 1 (poor) to 10 (excellent). The qualitative scale may be determined by the system creator.

The categories, knowledge, skills and terms are preferably set up in a relational database prior to the extraction process. As described above with reference to Figures 4 and 5, in a preferred embodiment, the relationship between categories and knowledge is many-many, the relationship between knowledge and skills is many-to-many, and the relationship between skills and terms is one-to-many.

Referring now to Figure 8, there is shown a flow chart of a preferred embodiment of a method for the present invention. In a preferred embodiment, a resume is evaluated (802) for a particular skill. The skill level for that particular skill is then determined (804) using the above described techniques. After a final skill level value is determined, the skill level is mapped (806) to a qualitative scale. Finally, the skill and the qualitative scale value of the skill level is stored (808) in the target database. More specifically, the categories, knowledge, skills and terms (i.e. the semantic network) are loaded into main memory. The electronic document text is then passed to the skill and knowledge information extractor 702. In a preferred embodiment, knowledge, skills, skill levels and number of years are extracted from the electronic document in the following manner: first, all the terms in the database are

checked against the document, then an initial scan of the document collects all the terms. The frequency of appearance of the term is recorded. Afterwards, the weightage factors for the skill level calculation are applied. A second scan of the electronic document analyses the document and a running list is maintained for all terms to calculate the experience duration where the term is maintained. On completion of the second scan, all the terms are rolled up into skills according to the semantic network and thesaurus, all the skills are rolled up into knowledge according to the semantic network, and all the knowledge items are rolled up into categories. Additionally, categories specifically mentioned are added. Thus, based on this information, the skill levels and years of experience are computed as described above.

Referring now to Figure 9, there is shown a screen shot of a user interface of a preferred embodiment of a target database for a skill and knowledge information extractor 702. Window 902 displays the particular skills analyzed from a candidate's resume, the qualitative level determined by the skill and knowledge information extractor 702, and the years of experience the candidate has for the particular skill. For example, the highlighted portion of window 902 indicates that the candidate has some skill as an analyst, that the qualitative proficiency of the candidate's skill as an analyst is "excellent", and that the candidate has 4 years of experience as an analyst. Thus, the present invention advantageously allows a user to extract, determine, and display from a candidate's resume the proficiency of a particular skill of the candidate.

The present invention is designed as a set of Object Oriented Libraries and contains the following major Object Libraries:

Xpert Object Library	This library encapsulates the Xerox InXight APIs and provides basic building blocks of extractions. For example, the date Xpert Object can decide whether a given text contains a date or not.
DataBase Library	This library encapsulates the ODBC APIs to connect to the target database, as well as provides Objects for creation and manipulation of extracted records.
Document Filter Library	This library provides Objects that can filter the input document, decide the document type and formatting. For example, the Word Document Object can decide if the input document is a Microsoft Word document or not and can extract the text from the Word files.
Paragraph Property Library	This library provides Objects that provide mechanisms to gather the paragraphs from the input document, Heuristics Objects which gather the Paragraph Properties from the input document, etc.... This information is later used in the extraction process.
Extraction Object Library	This library mainly contains one Object for each of the sections typically present in a document. For example, for a Resume document, it has a Resume Object, Section Object, Experience, Education, HonorsandAwards, Publications, Patents, Objective, References, CoverLetter, and Contact Information Objects. Each of these Objects has all the logic necessary to extract content from that particular section in a Resume.
Knowledge and Skill Level Object Library	This Library provides objects and facilities for extractive Knowledge of a person from a resume and provides objects and functions to calculate the skill level of a particular skill using the procedure described in the previous sections.
Thread Library	This library provides Objects for multi-threading and preferably encapsulates the Win32 thread API. It provides mechanisms for thread synchronization, semaphores, lock and resource management.

In a preferred embodiment, the present invention may be implemented to run on a Windows NT Server and any relational database such as Oracle Database. Database tables may be used to define how information is represented in a relational or object-oriented database. In an object-oriented implementation, any relational table is

preferably represented as an object class. The following section describes a preferred embodiment of the content and type of the fields that are extracted into a relational database, and also the definitions of the categories, knowledge, skills and terms. The supporting tables are also explained. One skilled in the art will realize that these tables are not limited to the specific information illustrated therein but may be created as needed, depending on the document type being processed.

Table 1

AutoEntryDocuments

Table 1 holds the documents that are to be extracted. It holds the following information:

DocID	Document ID
AEDocFileName	The complete path and file name of the document
AerfCategory	Category of the Resume/Document
Task_ID	TaskID associated with the document
AEStatus	Status of Extraction. (Not Done, Done, Errors)

Table 2

AutoEntrySchedule

Table 2 holds information about the scheduled extraction tasks.

Task_ID	Task ID of the Schedule
AEScheduleDate	Date the Task is Scheduled to Run
AESrfCategory	Category of the documents scheduled for this Task
AEActualDate	Date the Task is scheduled to Run
. . AEScheduleType	Type of the Schedule (Daily, Weekly, Monthly)
AEScheduleStatus	Status of the Task (Scheduled, Completed, Errors)

Table 3

Candidate

Table 3 holds the personal information like name of the person, contact address, current employer, resume summary etc. The XtractionXpert automatically extracts the following information from the resume:

CAFirstName	First Name of the Person
CALastName	Last Name of the Person
CANickName	Nick Name of the Person
CAWorkCompany	Current Employer
CATitle	Current Designation

CAYearsEmployed	Total Number of years of experience
CandidateID	Candidate ID, the database ID of the person
CASalutation	Salutation (Mr, Ms, Dr etc.)
CACurrentObjective	Stated Objective in the Resume
CABriefExperience	Text of the Summary Section
CAYearsOfExperienc	Years of Experience with current employer
CALastModifiedDate	Date of Last Modification to the Record
CATextResume	Actual text of the resume
CAModifiedBy	Person who modified the record
CAWorkMailStop	Mail Stop of the Work Address
CAWorkPhoneNo	Work Phone Number
CAWorkExtension	Work Phone Number Extension
CAWorkFaxNo	Work Fax Number
CAWorkMobilePhone	Work Related Mobile Phone Number
CAWorkEmail	Official Email Address
CAHomeMailStop	Mail Stop of Residence
CAHomePhoneNo	Home Phone Number
CAHomeExtension	Home Phone Extension
CAHomeFaxNo	Home Fax Number
CAHomeEmail	Home Email Address
CAOtherMailStop	Mail Stop other than Office and Residence
CAOtherPhoneNo	Any other Phone Number (e.g. Recruiting Agency)
CAOtherExtension	Extension number of Phone
CAOtherFaxNo	Fax Number other than work and residence
CAOtherMobilePhone	Mobile Phone Number
CAOtherEmail	Email Address other than residence and work
CAWorkStreet	Street Name of the Work address
CAWorkSuiteNo	Suite Number of Work address
CAWorkCity	City Name of the Work Address
CAWorkState	State Name of the Work Address
CAWorkZip	Zip Code of the Work Address
CAWorkCountry	Country of the Work Address
CAHomeStreet	Street Name of the Home Address
CAHomeSuiteNo	Suite/Apt. Number of the Home
CAHomeCity	City of Residence Address
CAHomeState	State of Residence Address
CAHomeZip	Zip Code of the Residence Address
CAHomeCountry	Country of the Residence Address
CAOtherStreet	Street Name of address other than work and home
CAOtherSuiteNo	Suite or Apt. Number other than work and phone
CAOtherCity	Name of City from the other address
CAOtherState	State Name of the other address

CAOtherZip	Zip Code of the other address
CAOtherCountry	Country of the other address
CAHomeMobilePhone	Mobile phone of other address
CAHomePage	Web address
CAPager	Pager Number

Table 4

ExperienceDetail

CandidateID	Database ID of the Person (Candidate Table)
EDEmployerName	Name of the Company worked for
EDReportedTo	Name of the reporting Manager
EDResponsibilityLl	Primary Responsibility (Designation)
EDResponsibilityL2	Secondary Responsibility
EDPeopleManaged	Number of people managed
. . EDHighlightsl	First Bulleted item from the Experience Description
EDHighlights2	Second Bulleted item
EDHighlights3	Third Bulleted Item
EDHighlights4	Fourth Bulleted Item
EDNotes Text of the	Experience Description
ExperienceDetailID	ID ofthe Record
EDStartDateDD	Date Joined for the company
EDStartDateMM	Month joined for the company
EDStartDateYYYY	Year joined for the company
EDEndDateDD	Date last worked for the company
EDEndDateMM	Month last worked for the company
EDEndDateYYYY	Year last worked for the company
EDReportedToPhone	Manager's Phone Number

5

Table 5

EducationRecord

CandidateID	Database ID of the person
EdrfMajor	Specialization
EDAwardedby	Name of the Institution
EDGPA	GPA earned
EDNote	Text of the description of the record
EdrfGradStatus	Status of graduation (passed, pending etc.)
EdrfDegreeType	Type of the Degree (B.S., M.S, Ph.D)
EDStartDateDD	Date joined in the course
EDStartDateMM	Month of joining
EDStartDateYYYY	Year of joining

EDEndDateDD	Date of Completion
EDEndDateMM	Month of Completion
EDEndDateYYYY	Year of Completion

Table 6

PatentRecord

CandidateID	Database Id of the Person
PATitle	Title of the Patent
Pacountry	Country where Patent was filed
PAJointHolder	Name of the Joint Holder
PAPatentNumber	Patent Number
PAGrantDateYYYY	Year Patent Granted
PAPatentStatus	Status of the Patent (Granted, Pending)
PANotes	Text of the description of the Patent
PAGrantDateMM	Month Patent granted
PAGrantDateDD	Date Patent granted

Table 7

PublicationRecord

CandidateID	Database Id of the Person
PurfPublicatType	Type of Publication (Book, Paper etc.)
PUTitle	Title of Publication
PUPublicationName	Name of the Publication
PUDateDD	Date of Publication
PUPublisherName	Name of the Publisher
PUDateMM	Month of Publication
PUPageRange	Page Numbers
PUDateYYYY	Year of Publication
Puisbn	ISBN number of the Publication
PUNotes	Text of the description

Table 8

SkillRecord

CandidateID	Database Id of the Person
ConceptID	Pointer to Concept Table where Skill Name is found
SKEXpYears	Number of years of experience in the skill
SKNotes	description of the skill
SkrfProfLevel	Skill Level

Table 9
Kno wledgeRecord

CandidateID	Database ID of the Person
KNYear	Number of years of experience
MetaConceptID	Pointer to MetaConcept
KNComment	Any comments associated

Table 10
ProfAssocRecord

CandidateID	Database ID of the person
PRAssocName	Name of the Professional Body
PRMemberCategory	Type of the membership
PRAssociatedSinceDD	Date of joining
PRAssociatedSinceMM	Month of joining
PRAssociatedSinceYYYY	Year of joining

Table 11
ProfLicenseRecord

CandidateID	Database ID of the person
PLLicenceName	Name of the License
PLLicenceAuthority	Name of the organization that issued the license
PLLicenceNumber	Professional License Number
PLLicenceGranted	Whether License Granted
PLLicenceLevel	Level of the License
PLLicenceState	Current status of the license
PLNotes	Text of the description
PLExpirationDateDD	Expiration Date (Date)
PLExpirationDateMM	Expiration Date (Month)
PLExpirationDateYYYY	Expiration Date (Year)

Table 12
ReferenceRecord

RECandidateRole	Role Played by the Candidate in the Team
CandidateID	Database ID of the Candidate
RERefrenceName	Name of the Referee
REReferenceTitle	Title (Designation) of the Referee
REWorkPhoneNo	Work Phone Number of the Referee
REHomePhoneNo	Home Phone Number of the Referee
RECandidateRelation	Relationship of the Referee to the Candidate
RECandidateDateBegin	Date candidate started the associationship

RECandidateDate End	Date association ended

Table 13

Courses

COCourseName	Name of the Course taken
CandidateID	Database Id of the Candidate
CODDateDD	Date course taken (date)
CODateMM	Date course taken (month)
CODateYYYY	Date course taken (year)
CoNotes	Description of the course

Table 14

AwardsHonors

Awhighlight	Name and highlight of the Award or Honor
CandidateID	Database Id of the candidate
AWNotes	Description of the Award or Honor

Table 1

MiscellenousInformation

MINotes	Text of the any other section
CandidateID	Database Id of the candidate

Table 16

Category

Table 16 provides information regarding the relationships between categories and knowledge information.

MetaConceptID	Knowledge ID
CmrfCategory	Category ID

Table 17

MetaConcept

Table 17 provides knowledge information for semantic network 220.

MetaConceptID	Knowledge ID
MEMetaConceptName	Name of the Knowledge Entry
Medescription	Description of the Knowledge Entry
MESemMarer	Semantic Markers and Types

Table 18

Concept

Table 18 provides information relating to skills.

ConceptID	Skill ID
CNConceptName	Name of the Skill
CNDescription	Description of the Skill
CNSemMarer	Semantic Markers and Types

Table 19

ConceptRelation

Table 19 provides information on relationships between skills and knowledge.

MetaConceptID	Knowledge ID
ConceptID	Skill ID
CRRelationType	Type of the Relation between Knowledge and Skill
CrisaRelationYN	Specifies if the relation is hierarchial
CRDescription	Description of the relation

Table 20

Term

Table 20 provides information on terms.

TermID	Term ID
TETerm	Name of the Term
LanguageID	Language ID of the Term
ConceptID	Skill ID to which the Term belongs

Table 21

Language

Table 21 stores information about different languages to which the terms belong.

LanguageID	Language ID
LALanguage	Name of the Language (English, French etc.)

Table 22

CaWord

CWWordID	Word ID
CWClassification	Classification of the word
CWWord	Word found

Table 23
CaWordList

CWWordID	Word ID
CWLFirstDocNumber	The First Document in which the word was found
CWLBlock	Block which consists the document Ids
CWLFlag	Database Flag

Table 24
CaWordPosition

CWWordID	Word ID
CWPFistDocNumber	First Document in which the word was found
CWPFlag	Database Flag
CWPBlock	Block which consists of document ids

From the above description, it will be apparent that the invention disclosed herein provides a novel and advantageous system and method for extracting and analyzing skill and knowledge information from an electronic document. The foregoing discussion discloses and describes merely exemplary methods and embodiments of the present invention. As will be understood by those familiar with the art, the invention may be embodied in other specific forms without departing from the spirit of the invention or essential characteristics thereof. Accordingly, the disclosure of the present invention is intended to be illustrative, but not limiting, of the scope of the invention, which is set forth in the following claims.

CLAIMS

1. An apparatus for extracting skill and knowledge information from an electronic document and for storing skill and knowledge information into a target database, the apparatus comprising:

a content analysis and semantic network engine for analyzing and extracting skill and knowledge information from the electronic document; and

a skill and knowledge information extractor, coupled to the content analysis and semantic network engine, for determining a skill level for the skill information extracted from the electronic document and for storing the skill level in the target database.

2. The apparatus of claim 1 wherein the skill and knowledge information extractor also maps the skill level for the skill to a qualitative scale.

3. The apparatus of claim 1 wherein the content analysis and semantic network engine further comprises:

a thesaurus for linking together terms and skills; and

a semantic network, coupled to the thesaurus, for organizing terms and skills of the thesaurus, knowledge, and categories, and for defining relationships between and among the terms, skills, knowledge, and categories.

4. The apparatus of claim 1 wherein the skill and knowledge information extractor determines a skill level, at least in part, by using the mathematical equation:

$$SkillLevel(X) = SkillLevel(X) + W(O) + \sum W(P_i) + W(K) + W(Y) - W(LU)$$

5. The apparatus of claim 1 wherein the skill and knowledge information extractor determines a skill level using a non-monotonic and default reasoning approach.

6. The apparatus of claim 2 wherein a skill extracted from the electronic document and the skill mapping to a qualitative scale are displayed on a computer.

7. An apparatus for analyzing and extracting skill and knowledge information from an electronic document into a target database having predefined fields, the apparatus comprising:

a thesaurus for linking together terms and skills and for defining relationships between and among the terms and skills; and

a semantic network coupled to the thesaurus for organizing terms and skills in the thesaurus, knowledge, and categories in a hierarchical structure;

wherein the thesaurus and semantic network are used to analyze skill and knowledge information in the electronic document.

8. The apparatus of claim 7 further comprising:

a document pre-processor coupled to the semantic network for classifying the electronic document as a document type and for performing an initial analysis on the electronic document.

9. The apparatus of claim 7 further comprising:

a heuristics engine coupled to the semantic network for applying a set of heuristics to the electronic document.

10. The apparatus of claim 7 further comprising:

a skill and knowledge information extractor for extracting skill and knowledge information from the electronic document and for determining a skill level for skill information extracted from the electronic document.

11. The apparatus of claim 10 further comprising:

a target database coupled to the semantic network for storing skill and skill level information in predefined fields in the target database.

12. A method for determining a skill level for skill information extracted from an electronic document, the method comprising the steps of:

identifying skill and knowledge information in the electronic document;

extracting the skill and knowledge information from the electronic document; and

determining a skill level for skill information extracted from the electronic document.

13. The method of claim 12 wherein the step of determining a skill level is performed by a skill and knowledge information extractor.

14. The method of claim 12 wherein the step of identifying skill and knowledge information is performed using a semantic network.

15. A method for processing skill and knowledge information from an electronic document, the method comprising the steps of:

identifying skill and knowledge information in the electronic document;

extracting the skill and knowledge information from the electronic document;

determining a skill level for skill information extracted from the electronic document; and

mapping the skill level to a qualitative scale.

16. A computer implemented method for extracting and displaying skill and knowledge information from an electronic document, the method comprising the steps of:

identifying skill and knowledge information in the electronic document;

extracting the skill and knowledge information from the electronic document;

determining a skill level for skill information extracted from the electronic document; and

mapping the skill level to a qualitative scale.

17. A computer-readable medium for extracting and displaying skill and knowledge information from an electronic document, the computer-readable medium comprising code for performing the steps of:

identifying skill and knowledge information in the electronic document;

extracting the skill and knowledge information from the electronic document;

determining a skill level for skill information extracted from the electronic document; and

mapping the skill level to a qualitative scale.

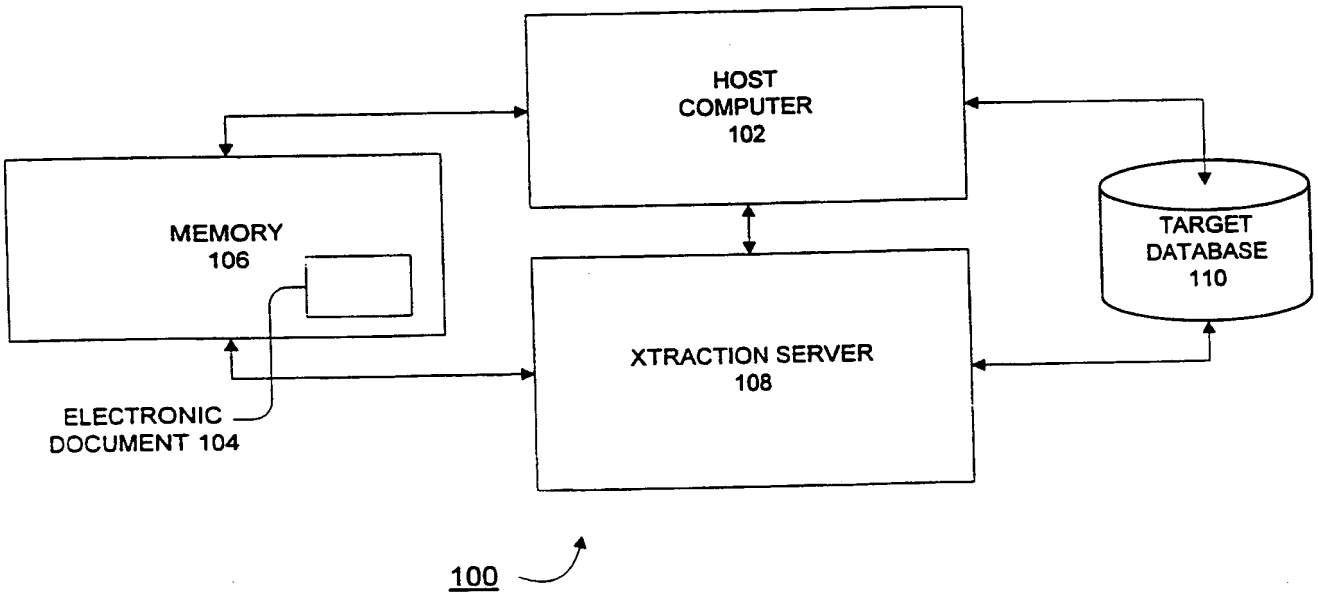

HOST
COMPUTER
102
MEMORY
106
TARGET
DATABASE
110
XTRACTION SERVER
108
ELECTRONIC
DOCUMENT 104
100

FIGURE 1

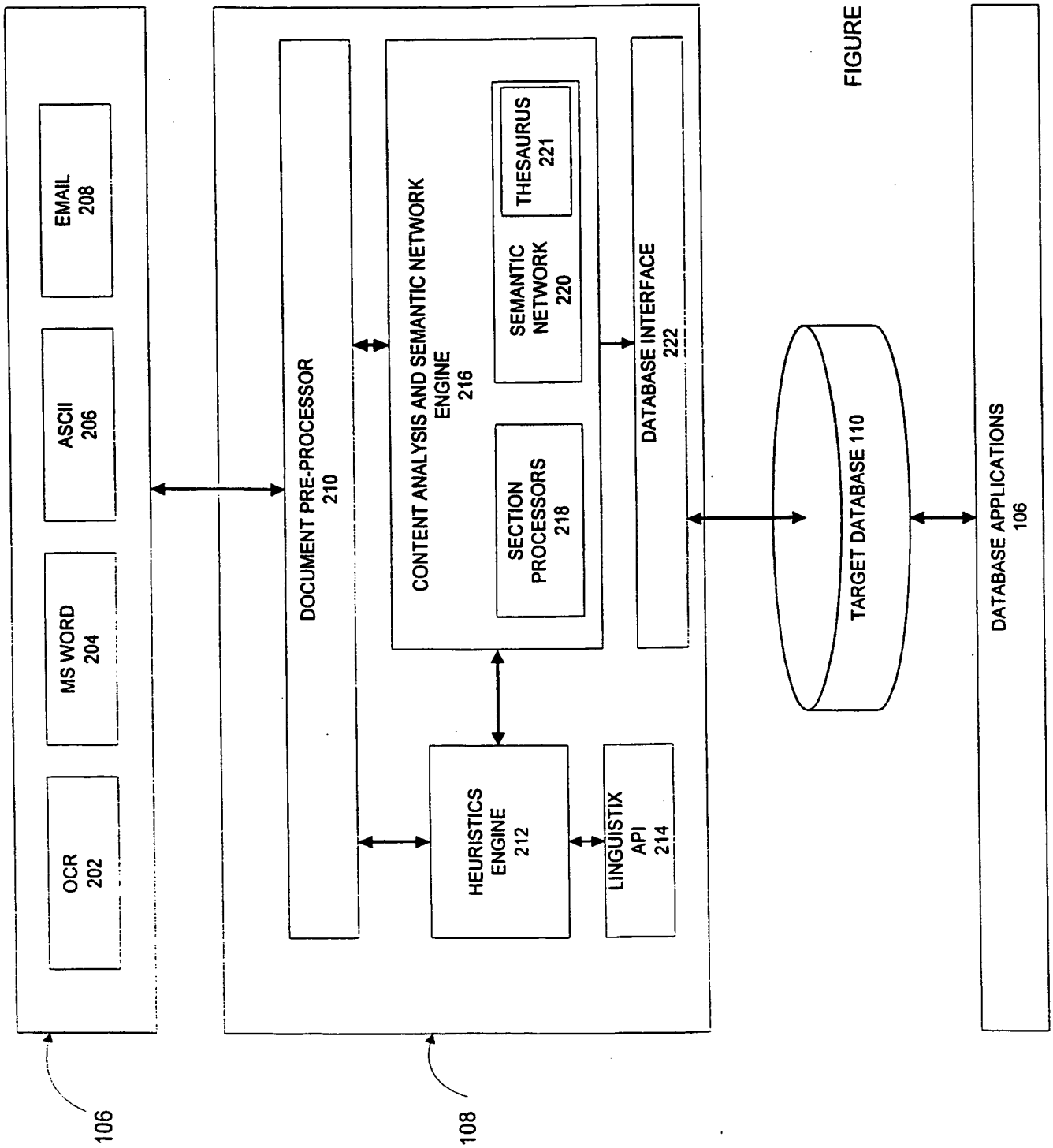

OCR 202
MS WORD 204
ASCII 206
EMAIL 208
106
DOCUMENT PRE-PROCESSOR 210
CONTENT ANALYSIS AND SEMANTIC NETWORK ENGINE 216
SECTION PROCESSORS 218
SEMANTIC NETWORK 220
THESAURUS 221
DATABASE INTERFACE 222
HEURISTICS ENGINE 212
LINGUISTIX API 214
108
TARGET DATABASE 110
DATABASE APPLICATIONS 106

FIGURE 2

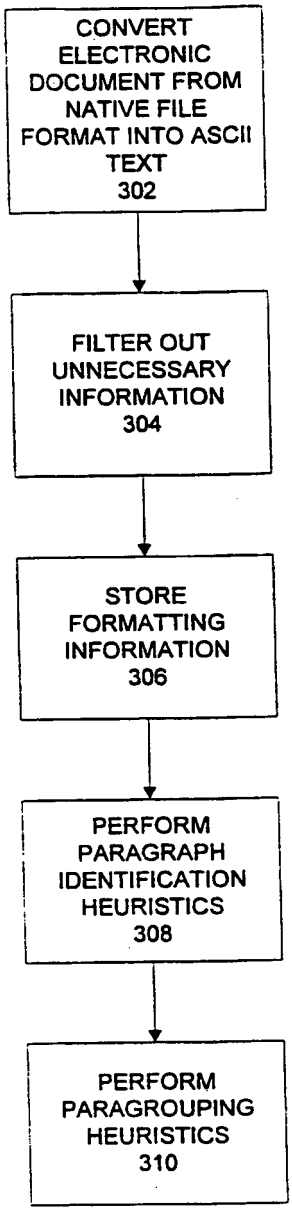

CONVERT ELECTRONIC DOCUMENT FROM NATIVE FILE FORMAT INTO ASCII TEXT 302
FILTER OUT UNNECESSARY INFORMATION 304
STORE FORMATTING INFORMATION 306
PERFORM PARAGRAPH IDENTIFICATION HEURISTICS 308
PERFORM PARAGROUPING HEURISTICS 310

FIGURE 3

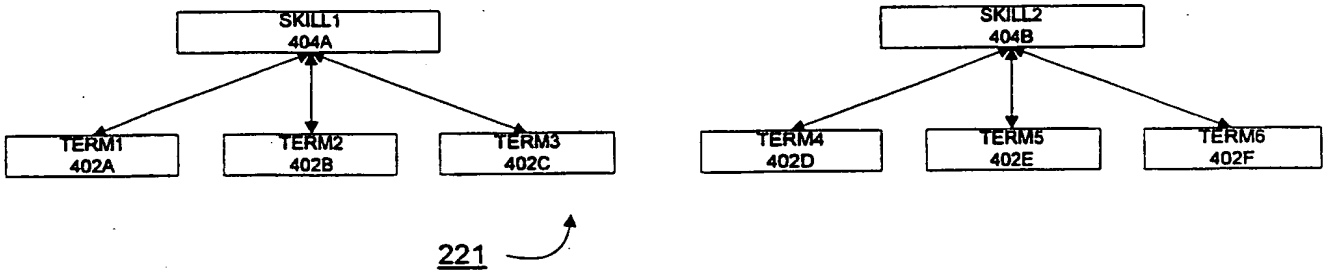


FIGURE 4

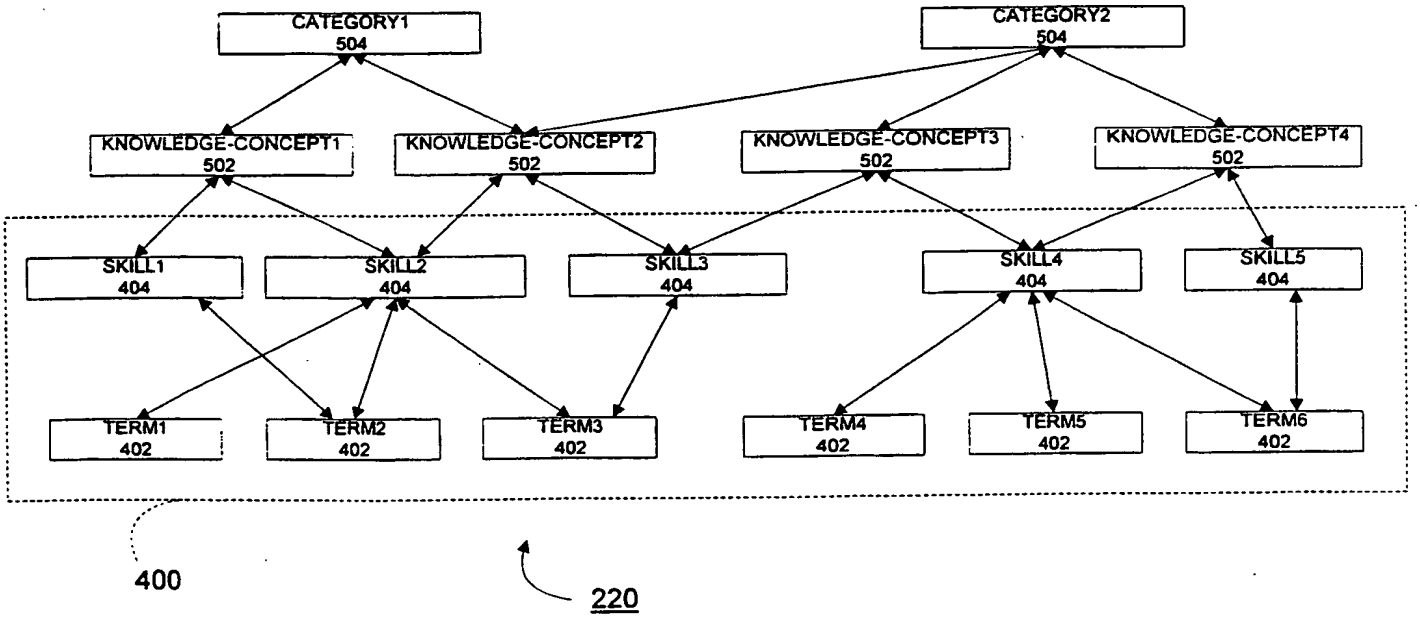


FIGURE 5

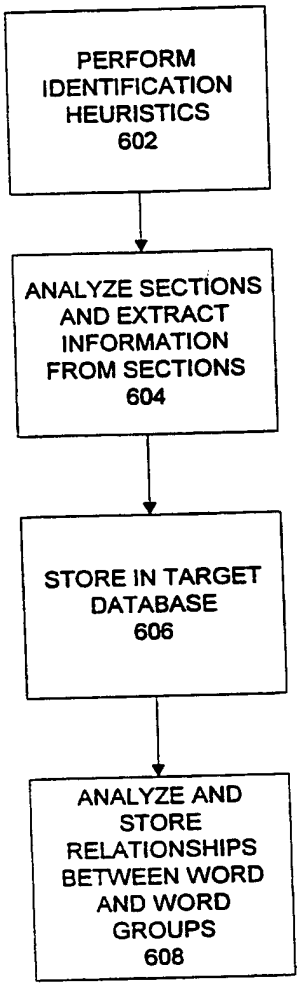

PERFORM IDENTIFICATION HEURISTICS 602
ANALYZE SECTIONS AND EXTRACT INFORMATION FROM SECTIONS 604
STORE IN TARGET DATABASE 606
ANALYZE AND STORE RELATIONSHIPS BETWEEN WORD AND WORD GROUPS 608

FIGURE 6

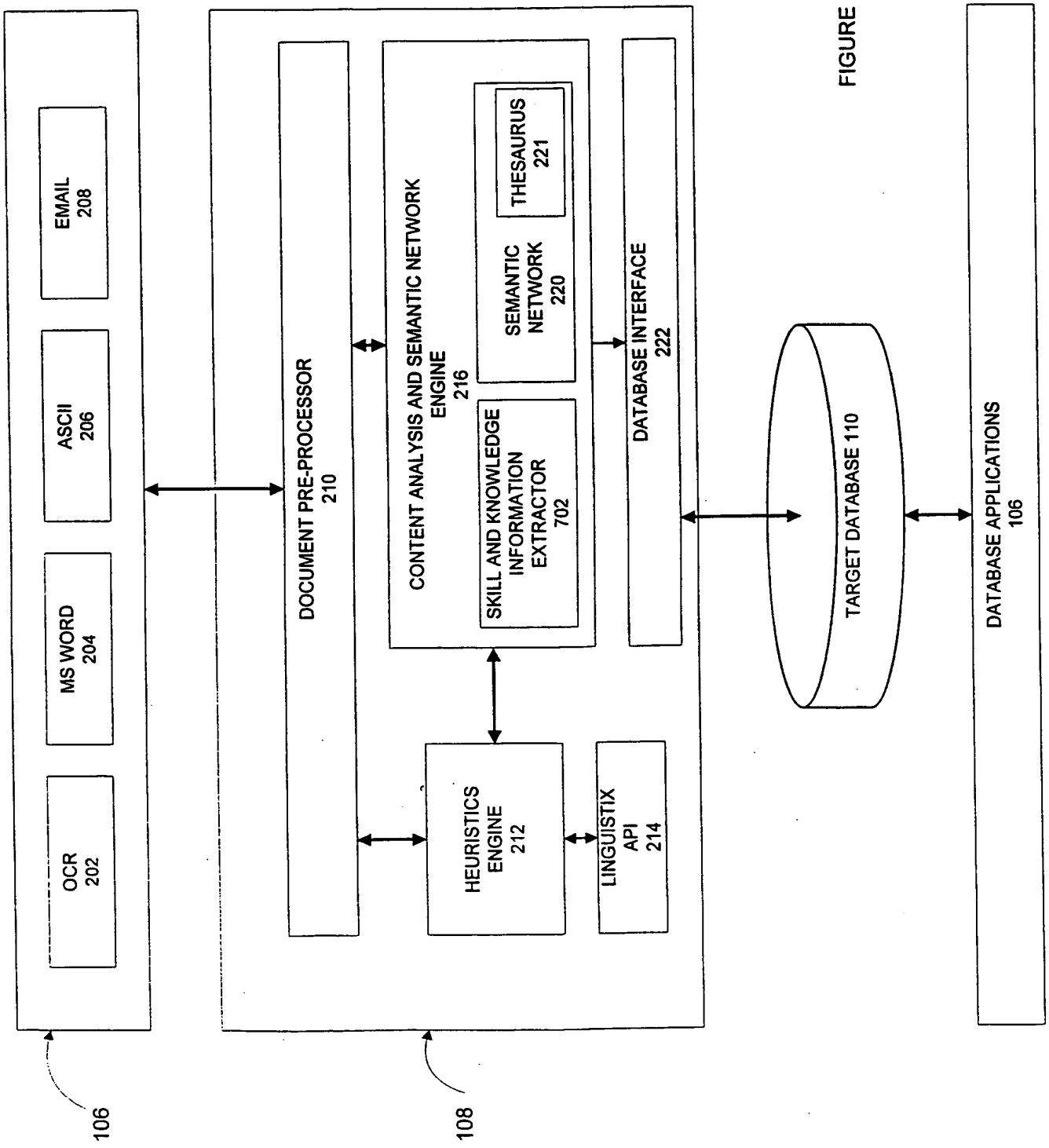

OCR 202
MS WORD 204
ASCII 206
EMAIL 208
106
DOCUMENT PRE-PROCESSOR 210
CONTENT ANALYSIS AND SEMANTIC NETWORK ENGINE 216
SKILL AND KNOWLEDGE INFORMATION EXTRACTOR 702
SEMANTIC NETWORK 220
THESAURUS 221
HEURISTICS ENGINE 212
LINGUISTIX API 214
108
DATABASE INTERFACE 222
TARGET DATABASE 110
DATABASE APPLICATIONS 106

FIGURE 7

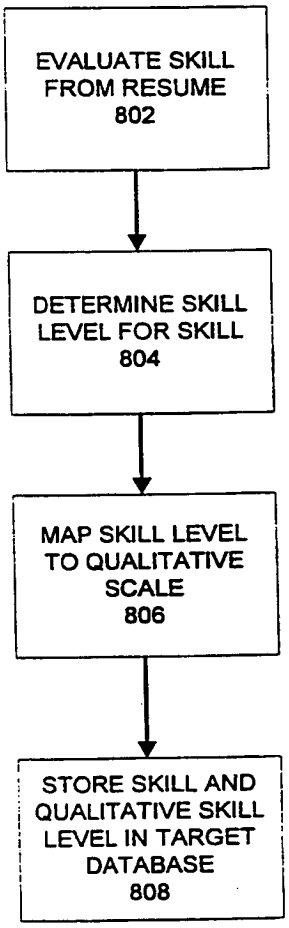

EVALUATE SKILL FROM RESUME 802
DETERMINE SKILL LEVEL FOR SKILL 804
MAP SKILL LEVEL TO QUALITATIVE SCALE 806
STORE SKILL AND QUALITATIVE SKILL LEVEL IN TARGET DATABASE 808

FIGURE 8

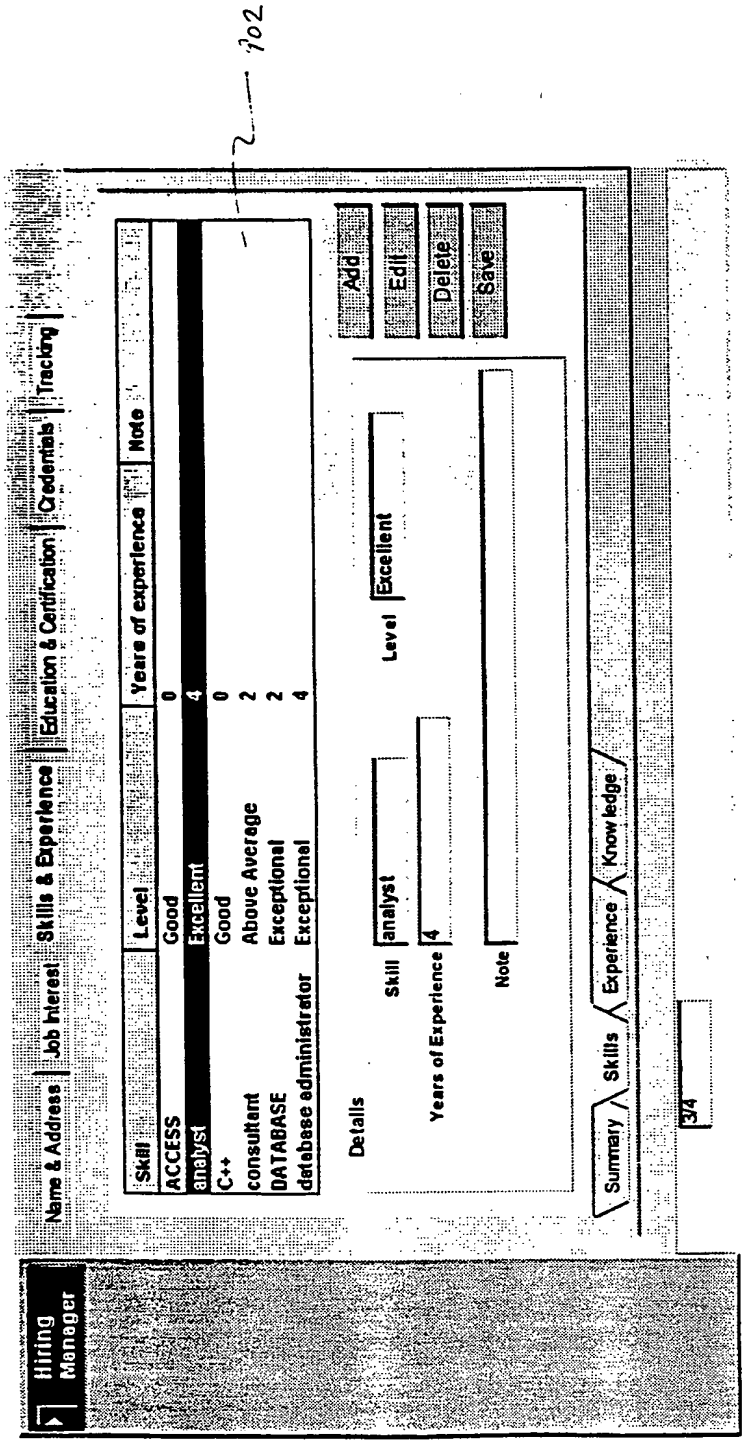


FIGURE 9

INTERNATIONAL SEARCH REPORT

:national Application No
PCT/US 99/26083

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/60

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 197 004 A (SOBOTKA DAVID ET AL) 23 March 1993 (1993-03-23) abstract column 3, line 38 -column 3, line 59 column 4, line 22 -column 4, line 68 column 5, line 49 -column 6, line 40 column 7, line 10 -column 8, line 58 claims figures 1,2,5-8	1-17
Y	US 5 297 039 A (KANAEGAMI ATSUSHI ET AL) 22 March 1994 (1994-03-22) abstract column 2, line 7 -column 6, line 59 figures 3,5,6	1-17

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
31 March 2000	07/04/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Abbing, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

rnational Application No
PCT/US 99/26083

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document. with indication.where appropriate. of the relevant passages	Relevant to claim No.
A	WO 98 39716 A (ELECTRONIC DATA SYST CORP) 11 September 1998 (1998-09-11) abstract page 9, line 5 -page 11, line 7	1-3,6-17
A	US 5 416 694 A (PARRISH EVE J ET AL) 16 May 1995 (1995-05-16) abstract column 1, line 30 -column 2, line 23 column 3, line 52 -column 4, line 24 column 5, line 1 -column 6, line 38	1-3,6-17
A	NESTOROV S ET AL: "Inferring structure in semistructured data" SIGMOD RECORD,US,SIGMOD, NEW YORK, NY, vol. 26, no. 4, May 1997 (1997-05), pages 39-45-43, XP002099175 ISSN: 0163-5808 the whole document	1-3,6-17

1

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

national Application No
PCT/US 99/26083

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5197004 A	23-03-1993	NONE	
US 5297039 A	22-03-1994	JP 2943447 B	30-08-1999
		JP 4357568 A	10-12-1992
WO 9839716 A	11-09-1998	AU 6182098 A	22-09-1998
US 5416694 A	16-05-1995	CA 2140216 A	29-08-1995

CORRECTED VERSION*

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/60

A1

(11) International Publication Number: WO 00/26839

(43) International Publication Date: 11 May 2000 (11.05.00)

(21) International Application Number: PCT/US99/26083

(22) International Filing Date: 3 November 1999 (03.11.99)

(30) Priority Data:

60/107,063	4 November 1998 (04.11.98)	US
PCT/US98/27664	28 December 1998 (28.12.98)	US
09/380,219	27 August 1999 (27.08.99)	US

(71) Applicant *(for all designated States except US)*: INFODREAM CORPORATION [–/US]; 2340A Walsh Avenue, Santa Clara, CA 95051 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: ANDLEIGH, Prabhat, K. [US/US]; 10701 Castine Avenue, Cupertino, CA 95014 (US). PAPPU, Nagaraju [IN/IN]; Ashraya Residency, Apt. No. 54, 27th Main, 3rd Cross, BTM Layout – I Phase, Bangalore 560068 (IN). KALINDINDI, Vasudeva, V. [IN/IN]; Apartment 95, 3655 Pruneridge Avenue, Cupertino, CA 95051 (US).

(74) Agents: RADLO, Edward, J. et al.; Fenwick & West LLP, Two Palo Alto Square, Palo Alto, CA 94306 (US).

(81) Designated States: CA, GB, IN, US.

Published

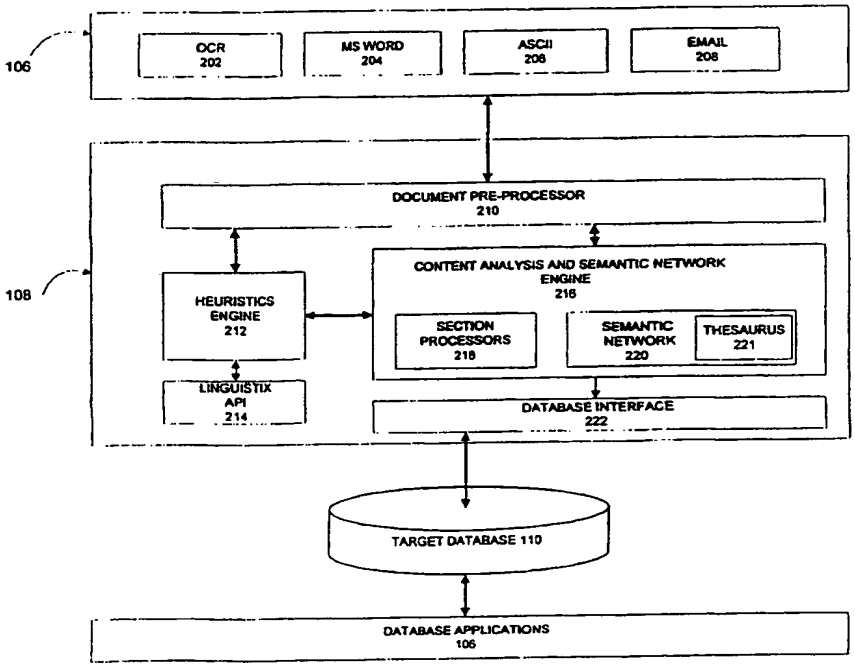
With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: ADVANCED MODEL FOR AUTOMATIC EXTRACTION OF SKILL AND KNOWLEDGE INFORMATION FROM AN ELECTRONIC DOCUMENT

(57) Abstract

An apparatus, method, and computer readable medium for analyzing and extracting skill and knowledge information from an electronic document (104) and for storing the extracted skill and knowledge information into predefined fields or tables in a target database (110) comprises a content analysis and semantic network engine (216) for analyzing and extracting skill and knowledge information from the electronic document (104). A skill and knowledge information extractor (702) is coupled to the content analysis and semantic network engine (216), for determining a skill level for the skill information extracted from the electronic document (104). In a preferred embodiment, the skill and knowledge section processor (702) uses a non–monotonic reasoning principle to determine a skill level for skill information extracted from the electronic document (104). The content analysis and semantic network engine (216) further comprises a thesaurus (221) for linking together terms (402) and skill information (404), and for defining relationships between and among the terms (402) and skill information (404), and a semantic network (220) coupled to the thesaurus (221), for organizing the terms (402) and skill information (404) in the thesaurus (221), along with knowledge information (502) and categories (504), in a hierarchical structure.



*(Referred to in PCT Gazette No. 41/2000, Section II)

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 May 2000 (11.05.2000)

PCT

(10) International Publication Number
WO 00/26839 A1

(51) International Patent Classification[7]: G06F 17/60

(21) International Application Number: PCT/US99/26083

(22) International Filing Date:
3 November 1999 (03.11.1999)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/107,063 4 November 1998 (04.11.1998) US
PCT/US98/27664
28 December 1998 (28.12.1998) US
09/380,219 27 August 1999 (27.08.1999) US

(71) Applicant *(for all designated States except US)*: **INFODREAM CORPORATION** [—/US]; 2340A Walsh Avenue, Santa Clara, CA 95051 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **ANDLEIGH, Prabhat, K.** [US/US]; 10701 Castine Avenue, Cupertino, CA 95014 (US). **PAPPU, Nagaraju** [IN/IN]; Ashraya Residency, Apt. No: S4, 27th Main, 3rd Cross, BTM Layout - I Phase, Bangalore 560 068 (IN). **KALINDINDI, Vasudeva, V.** [IN/IN]; Apartment 95, 3655 Pruneridge Avenue, Cupertino, CA 95051 (US).

(74) Agents: **RADLO, Edward, J.** et al.; Fenwick & West LLP, Two Palo Alto Square, Palo Alto, CA 94306 (US).

(81) Designated States *(national)*: CA, GB, IN, US.

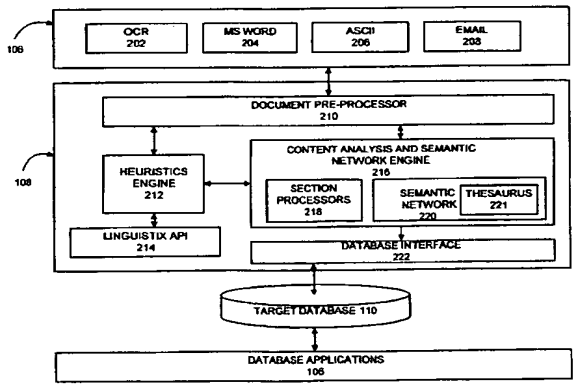
Published:
— *with international search report*

(48) Date of publication of this corrected version:
2 August 2001

(15) Information about Corrections:
see PCT Gazette No. 31/2001 of 2 August 2001, Section II
Previous Correction:
see PCT Gazette No. 41/2000 of 12 October 2000, Section II

[Continued on next page]

(54) Title: ADVANCED MODEL FOR AUTOMATIC EXTRACTION OF SKILL AND KNOWLEDGE INFORMATION FROM AN ELECTRONIC DOCUMENT



(57) Abstract: An apparatus, method, and computer readable medium for analyzing and extracting skill and knowledge information from an electronic document (104) and for storing the extracted skill and knowledge information into predefined fields or tables in a target database (110) comprises a content analysis and semantic network engine (216) for analyzing and extracting skill and knowledge information from the electronic document (104). A skill and knowledge information extractor (702) is coupled to the content analysis and semantic network engine (216), for determining a skill level for the skill information extracted from the electronic document (104). In a preferred embodiment, the skill and knowledge section processor (702) uses a non-monotonic reasoning principle to determine a skill level for skill information extracted from the electronic document (104). The content analysis and semantic network engine (216) further comprises a thesaurus (221) for linking together terms (402) and skill information (404), and for defining relationships between and among the terms (402) and skill information (404), and a semantic network (220) coupled to the thesaurus (221), for organizing the terms (402) and skill information (404) in the thesaurus (221), along with knowledge information (502) and categories (504), in a hierarchical structure.

WO 00/26839 A1

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

ADVANCED MODEL FOR AUTOMATIC EXTRACTION OF SKILL AND KNOWLEDGE INFORMATION FROM AN ELECTRONIC DOCUMENT

RELATED APPLICATION

The subject matter of this application is a continuing application of and claims priority from U.S. patent application Serial No. 09/380,219, filed August 27, 1999 descending in priority from PCT application PCT/US98/27664, filed on December 28, 1998, and entitled "Xtraction Server" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi. Said two earlier applications are commonly assigned with the instant application.

The subject matter of this application is also related to and claims priority from U.S. Provisional Application Serial No. 60/107,063, filed Novmeber 4, 1998, and entitled "Advanced Model for Automatic Extraction of Content, Skills, and Knowledge from Resumes" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi, which application is commonly assigned with the instant application, and is incorporated herein by reference in its entirety.

TECHNICAL FIELD

This invention relates to the field of computer analysis of electronic documents. More specifically, it relates to the field of information retrieval to convert and store information in documents written in a natural language into a predefined structure which can be retrieved and manipulated by computer program applications.

BACKGROUND OF THE INVENTION

Information to be sorted and stored in a computer database may reside in numerous electronic documents. For example, information about people and their specific talents and skills may reside in electronic documents, such as resumes, performance appraisals, design documents, publications, books, patent documents, and email messages. When an individual

is trying to organize and sort out specific information from such electronic documents, the individual usually has to open each document separately and manually analyze, retrieve, and store the relevant data in the particular database. For example, a project manager who would like to find the best employee for a specific job may have a specific job description. When searching for an employee whose skills, knowledge and talent are best suited for the specific job description, the project manager must sift through several documents which contain the necessary information. Such a process is time consuming and inefficient, because the project manager may have to read the documents several times and may have to review and type the information into a computer database in order to organize the various pieces of information into a coherent summary.

A computerized system which can analyze and extract pertinent information from different electronic documents would provide a more efficient solution to this problem. However, such text documents are often written in unstructured natural language text for other people to understand. Thus, computer programs such as database applications cannot efficiently process documents written in natural language texts. Rather, computer programs can process only information which has been stored in a highly structured fashion in order to retrieve and manipulate that information. Additionally, these documents may be prepared in a variety of different file formats, such as Microsoft Word 97, Rich Text Format, PDF, WordPerfect, ASCII files, and HTML, and may be stored in different areas within a computer.

There are a variety of information retrieval programs such as Internet search engines that can retrieve documents that match a set of keywords. Their scope is very limited in the context of the above mentioned problem, because they cannot understand the text, and

certainly they cannot make any connection between the document and the person who is related to that document. Another problem is that the 'information of interest' will vary significantly from one organization to another. For example, a health care organization will be interested in the skills and talents related to the medical field, but the skills related to computers may not be of significant interest, whereas a software development organization will be interested in the computer and software related skills, but may not be interested in medical or first-aid related skills. The keyword based search engines cannot address this problem of retrieving only the 'information of interest'. As a result, there is a vast amount of information about people which cannot be easily processed by computer programs.

For example, in today's large corporations and government organizations, it is not uncommon to receive hundreds of thousands of resumes of potential candidates in a very short time. Recruiting the right candidates from such a vast pool of applicants is a very complicated problem. It is crucial for organizations to find the people with the right knowledge and skill set. In essence, managers have to deal with a vast number of resumes, try to understand the content within the resumes, and short-list candidates who have the right skills and knowledge. For example, if an organization wants to recruit a middle level manager with 5 to 8 years of experience to lead a development project, the organization will need to sort through thousands of resumes and determine from each one whether that particular candidate has the requisite knowledge and skill level. It is not possible to find the best resumes using a standard full text search engine because such search programs search for a particular input string and retrieve only resumes which contain that particular input string. Such an approach is not that useful, because a particular skill may be written using many different terms (e.g. Microsoft Word, MS Word, Word 97, etc....) even though the terms all

refer to the same or similar ideas. Moreover, in addition to not being able to correctly identify a candidate's skills, a typical search program cannot identify the type of experience with that skill, the duration of that experience, or the overall knowledge gained by the candidate in a specific skill group. Additionally, it is also very desirable to have a system for determining not only the knowledge and skills of a candidate but also the proficiency level of a candidate in a particular skill.

Therefore, what is needed is a system for analyzing and extracting information from an electronic document and for storing the extracted information in a database. Additionally, what is needed is a system for analyzing and extracting skill and knowledge information from an electronic document and for determining a skill level for skill information and for mapping such skill level information to a qualitative scale.

DISCLOSURE OF INVENTION

The present invention is an apparatus, method, and computer-readable medium for analyzing and extracting skill and knowledge information from an electronic document (104) and for storing the extracted skill and knowledge information into predefined fields or tables in a target database (110). The system for analyzing and extracting skill and knowledge information from an electronic document (104) comprises a content analysis and semantic network engine (216) for analyzing and extracting skill and knowledge information from the electronic document (104), and a skill and knowledge information extractor (702) coupled to the content analysis and semantic network engine (216), for determining a skill level for the skill information extracted from the electronic document (104). In a preferred embodiment, the skill and knowledge section processor (702) uses a non-monotonic reasoning principle to determine a skill level for skill information extracted from the electronic document (104).

The content analysis and semantic network engine (216) further comprises a thesaurus (221) for linking together terms (402) and skill information (404) and for defining relationships between and among the terms (402) and skill information (404), and a semantic network (220) coupled to the thesaurus (221), for organizing the terms (402) and skill information (404) in the thesaurus (221), knowledge information (502), and categories (504) in a hierarchical structure.

A method for extracting skill and knowledge information from an electronic document (104) comprises the steps of: identifying skill and knowledge information in the electronic document (802); determining a skill level for skill information from the electronic document (804); and mapping the skill level to a qualitative scale (806). The method further comprises the step of storing the skill information and qualitative skill level scale mapping in the target database (808).

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram of a preferred embodiment of a system 100 in accordance with the present invention.

Figure 2 is a block diagram of a preferred embodiment of an extraction server 108 in accordance with the present invention.

Figure 3 is a flow chart of a preferred embodiment of the steps performed by the document pre-processor 210.

Figure 4 is a block diagram of a preferred embodiment of a thesaurus. 221

Figure 5 is a block diagram of a preferred embodiment of a semantic network 220.

Figure 6 is a flow chart of a preferred embodiment of the steps performed by the extraction server 108.

Figure 7 is a block diagram of a preferred embodiment of a system 700 in accordance with the present invention.

Figure 8 is a flow chart of a preferred embodiment of the steps performed by the skill and knowledge information extractor 702.

Figure 9 is a screen shot of a user interface of a preferred embodiment of a target database 110 display for skill information.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring now to Figure 1, a system 100 upon which a preferred embodiment of the present invention operates is shown. A host computer 102, using the method and system described herein, operates upon an electronic document 104, derived from a text document which contains unstructured text. As used herein "unstructured text" refers to any document which has been written in a natural language such as English. Examples of documents containing unstructured text include, but are not limited to, a resume, performance appraisals, design documents, publications, books, patent documents, and email messages. In a preferred embodiment, the host computer 102 is a conventional computer having a keyboard and mouse for input (not shown), and a conventional memory 106 associated with host computer 102 for storing the electronic document 104. The electronic document 104 may be prepared in any electronic file format, such as Microsoft Word 97, Rich Text Format, PDF, WordPerfect, ASCII files, and HTML.

The electronic document 104 is processed by host computer 102 using the present invention. Specifically, host computer 102 uses extraction server 108 to analyze, retrieve and store words and word groups from the electronic document 104 into a predefined structure in target database 110. As used herein, the terms "words" and "word groups" are used to mean

any text that may be derived from document 104 including, but not limited to, individual words or numbers, phrases, whole sentences, and blocks of text. The extraction server 108 identifies the document type of the document 104 and determines which words and word groups are to be extracted from the document 104. The structure and operation of the extraction server 108 is described in more detail below with reference to Figures 2 through 6.

The target database 110 comprises predefined tables with predefined columns for storing the word and word groups extracted from the electronic document 104. In a preferred embodiment, a predefined table and predefined columns correspond to a particular document type. For example, if document 104 is a resume, then a predefined table for a document type called "resume" may have predefined columns such as "name and address", "education", and "skills and experience". As another example, if document 104 is a patent document, then a predefined table for a document type called "patent document" may have predefined columns such as "inventors", "company", "patent number", and "field of search". The predefined tables and columns in target database 110 are organized ahead of time, and one skilled in the art will realize that the present invention is not limited to a particular document type or a predefined table, but that many different compilations of predefined tables and columns may be stored in target database 110 within the scope of this invention. The words and word groups stored in the target database 110 can be stored in electronic form on any type of computer data storage device or they may be printed out in a hard-copy printed format.

The process of extraction performed by the extraction server 108 preferably uses a non-monotonic reasoning principle. As used herein, a "non-monotonic reasoning principle" refers to a process whereby at every stage during extraction, the extraction server 108 assumes a reasonable default value. That default value is modified as further information

becomes available. For example, a string '1987' is first assumed to be a number, and if further information to qualify the string to be a date is available (for example in this case, that the string is preceded by another string 'Jan'), then the assumption is changed. If again further information becomes available to negate the previous assumption, the assumption is changed again.

Thus, the present invention advantageously allows a user to extract skill and knowledge information from an electronic document directly into a database. More specifically, the present invention analyzes an electronic copy of a text document and extracts words and word groups relating to skill and knowledge information into a target database comprising predefined tables and columns associated with a particular document type. Moreover, the present invention operates upon electronic documents in any electronic file format. The extracted skill and knowledge information stored in the target database can then be retrieved and manipulated by other computer program applications.

Referring now to Figure 2, a block diagram of a preferred embodiment of the extraction server 108 is shown. The electronic document 104 may be any electronic file stored in memory 106 which is accessible by the extraction server 108. For example, the electronic document 104 may be an electronic form of a hard copy of a document converted using a conventional optical scanner and Optical Character Recognition (OCR) software 202, a Microsoft Word file 204, an ASCII text file 206 or an email attachment 208. The database applications which manipulate the extracted information in target database 110 are also preferably stored in memory 106. In a preferred embodiment, the extraction server 108 comprises a document preprocessor 210 coupled to the memory 106 where the electronic document 104 is stored, a heuristics engine 212 coupled to the document pre-processor 210, a

morphological analysis engine 214 coupled to the heuristics engine 212, a content analysis and semantic network engine 216 coupled to the document preprocessor 210, and a database interface 222 coupled to the content analysis and semantic network engine 216 and to the target database 110. The content analysis and semantic network engine 216 preferably comprises section processors 218 and a semantic network 220.

The document pre-processor 210 retrieves the electronic document 104 from memory 106 and performs the initial analysis of the electronic document 104. Referring now to Figure 3, a flowchart of the steps of a preferred operation of the document pre-processor 210 is shown. The document pre-processor 210 performs the initial analysis and extraction of the electronic document 104 by first converting (302) the electronic document 104 from its native file format into ASCII text. More specifically, the document pre-processor 210 identifies the file format of the electronic document 104 and extracts the ASCII text out the document 104. For example, if the electronic document 104 is a Microsoft Word file, then the document pre-processor 210 identifies the file by the Microsoft Word signature and uses the Microsoft Object Linking and Embedding Software Development Kit (Microsoft OLE 2.0 SDK) to extract text from the Microsoft Word File.

Next, the document pre-processor 210 filters out (304) any unnecessary and unwanted information such as, but not limited to, email headers, OCR headers, blank pages, and unwanted characters. Preferably, any information that is not part of the original document is treated as unnecessary information. For example, email headers, non-ASCII characters at the beginning or at the end of the file, extra blank lines and blank spaces are removed from the text. Additionally, if the text contains vertical tables, these tables are preferably converted into horizontal tables. If the text contains multiple columns, it is preferably converted into

single column. The document pre-processor 210 then stores (306) formatting information for the document 104 such as, but not limited to, the fonts used, font sizes, section tittles, and subsections.

The document pre-processor 210 then performs paragraph identification heuristics (308) on the electronic document 104. During this step, the beginning and end of each paragraph is identified, and the paragraph characteristics are gathered. As used herein, the phrase "paragraph characteristics" refers to the statistical properties of the paragraph. Paragraph characteristics include, but are not limited to, the number of words in the paragraph, the number of lines in the paragraph, the average number of words per line, whether any line has a bullet as the starting character, and whether there are any underlined sentences in the paragraph.

Finally, the document pre-processor 210 performs paragraph grouping heuristics (310) on the electronic document 104. Once the paragraphs have been identified, the document pre-processor 210 groups the paragraphs into sections. During this step, the paragraphs are grouped into sections based on the paragraph characteristics as well as using any section tittles that precede the paragraphs. Starting at the beginning of the electronic document 104, the first heading or section title is identified, and the following paragraphs until the next section title are grouped into one section. If no section titles are found, then using the paragraph characteristics, all the similar paragraphs are grouped into sections. Additionally, paragraphs that have same or similar characteristics are grouped together into sections.

The heuristic engine 212 applies a set of heuristics, that is a set of rules, to the electronic document 104 for analyzing information in the electronic document 104. The set of heuristics which are applied to the electronic document 104 are associated with a particular

document type. For example, if the document type is a "resume", then the set of heuristics associated with the document type "resume" is applied to the electronic document 104. Heuristics are described below in more detail in commonly assigned U.S. patent application Serial No. 09/380,219 entitled "Extraction Server" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi, which is incorporated herein by reference in its entirety.

The morphological analysis engine 214 is used for target language analysis and is preferably the LinguistiX 2.0 application programming interface (API) from InXight Corporation in Palo Alto, CA. The LinguistiX 2.0 API is a language neutral programming interface. In other words, the LinguistiX API can analyze documents in any language such as English, French or German. Because the heuristics engine 212 and the LinguistiX API are external to and separate from the document pre-processor 210 and the content analysis and semantic network engine 216, the present invention can extract information from documents in the English, French or German language, and any other languages which will be supported by the LinguistiX API in future.

Preferably, the Heuristics Engine 212 uses the following features provided by the LinguistiX API: tokenization, lexical analysis, tagging, and noun-phrase extraction. Before text from the electronic document 104 can be analyzed in terms of its linguistic roots and function, it must first be segmented into words, punctuation and idiomatic phrases. LinguistiX tokenization includes the ability to recognize multi-word constructs such as HTML tags. The lexical analysis feature identifies the grammatical features of a word in addition to its root forms. The tagging feature identifies the grammatical category of words by their context. The noun-phrase extraction identifies multi-word phrases in documents. LinguistiX phrase extraction technology enables software to work with these larger concepts

to provide improved information analysis and retrieval. For example, 'Windows Programming' will be identified as one phrase, instead of two distinct words Windows and Programming. This feature is used by the semantic network 220 to identify the multi-word noun phrases.

These features of the LinguistiX API are used to implement the heuristics. For example, by using the tagging feature, the extraction server 108 may discover that a particular word is a proper noun. Whether that word is the name of the person or the name of a company will depend on where the word occurred in a document. For example, if the word occurs in a contact information section of a document, then it may be the name of the person, or name of the street, city and so on. If the word occurs in an experience section of a document, and if it is followed by the name of a city and state, it may be a company name.

The database interface 222 is a set of APIs that provide a mechanism for retrieving and storing information to and from the target database 110. This is done in such a way that the underlying implementation of the target database 110 is hidden from the application using the database interface. Thus, the extraction server 108 can work with any industry standard relational database software such as Oracle or Microsoft SQL Server without having to change the software or its implementation. Additionally, the database interface 222 provides the following mechanisms: a method to connect to the target database, a method to maintain the connection to the database, a transaction model to maintain the consistency of the database, and various methods to retrieve, query, update, insert and delete information from the target database 110.

The content analyzer and semantic network engine 216 analyzes the content of the electronic document 104, extracts words and word groups from the document 104, and stores

the extracted information in the appropriate tables in the target database 110. In a preferred embodiment, the content analyzer and semantic network engine 216 comprises section processors 218 which extract information from a particular section of interest, and a semantic network 220. The semantic network 220 uses a thesaurus 221 and a phrase extraction process to identify the meta-concepts and categories in the electronic document 104 and extracts related words and word groups into the target database 110. In a preferred embodiment, the present invention may be implemented to run on a Windows NT Server and Oracle Database.

Referring now to Figure 4, a block diagram of a preferred embodiment of a thesaurus 221 is shown. The thesaurus 221 is a vocabulary database for the extraction server 108 and is organized by skills. The thesaurus 221 groups all related terms 402 in a language under a language independent concept 404. As used herein, a "term" 402 refers to all the individual words or word groups that belong to a particular language along with their alternatives. As used herein, a "concept" or "skill" 404 comprises a set of terms 402 that are language specific and alternatives to one another. However, the skill 404 itself is language independent. Skills 404 establish synonymous relationships among all terms 402 in the thesaurus 221 that have the same meaning. In other words, skills 404 connect all the different names for the same skill 404 that are known to the thesaurus 221 and specify certain characteristics for each name. Preferably, each skill 404 has a unique skill identifier (ConceptID). The Concept ID by itself has no intrinsic meaning. Each term 402 in each language in the thesaurus 221 has a unique term identifier. The same term 402 in different languages, for example, in English and Spanish, will have a different term identifier for each language.

To illustrate the relation between terms 402 and skills 404 consider an example in which term1 402A may consist of 'MS VC++', term2 402B may consist of 'Microsoft Visual

C++' and term3 402C may consist of 'MS Visual C++'. All these terms 402 are linked to the skill 404 'Visual C++'. In other words, if the electronic document 104 uses any of the words or word groups 'MS VC++', 'Microsoft Visual C++' or 'MS Visual C++', the thesaurus 221 allows the extraction server 108 to recognize the words or word groups as being linked to the skill1 404A 'Visual C++'. In another example, term4, term5 and term6 are respectively 'JDK 1.1', 'Symantec Café', and 'JDBC', and all these terms 402 are linked to the skill2 404B called 'Java'. Thus, if the electronic document 104 uses any of the words or word groups 'JDK 1.1', 'Symantec Café', and 'JDBC', the thesaurus 221 allows the extraction server 108 to recognize the word or word group as being linked to the skill2 404B 'Java'.

The thesaurus 221 may also comprise other information such as the attributes of a skill 404 or attributes of a term 402. Attributes provide additional information that helps to define the meaning of a skill 404 and explain how it may be used in a document. In other words, the different senses of a particular word or word groups are captured using the attributes.

In addition to the relationship between a skill 404 and a set of terms 402, the thesaurus 221 also comprises relationships among skills 404. Preferably, these relationships are non-subsumption relationships. As used herein, the term "non-subsumption" refers to relationships that include related skills, co-occurring skills and/or associated expressions. In other words, non-subsumption refers to relationships that are not based on subsumption. For example, C++ and Java are related, but neither subsumes the other. All these relationships among skills 404 indicate that the skills 404 linked together are not exactly similar but are associated with each other in different ways. One skilled in the art will realize that the terms and skills of the thesaurus 221 are not limited to the examples given herein but may contain

any number of terms and skills which have been predefined and stored in the thesaurus 221 prior to the processing of the electronic document 104. Thus, the thesaurus advantageously allows the present invention to link together terms and skills used in specific industries, disciplines, and technologies for which the thesaurus is being used, and preserves the meanings and hierarchical connections between those terms and skills. Additionally, the thesaurus facilitates the access to concept relationships and to term and skill attributes irrespective of the term used as a point of entry.

Referring now to Figure 5, a block diagram of a preferred embodiment of a semantic network 220 is shown. The semantic network 220 provides a way of arranging all the skills 404 at the lowest level and then builds a taxonomy or network of higher level knowledge-concepts and categories. The semantic network 220 comprises skills 404 at the lowest level, "knowledge" or knowledge-concepts 502 at a second level, and categories 504 at the highest level. The semantic network 220 together with the thesaurus 221 provides a four level hierarchy of terms 402, skills 404, knowledge-concepts 502 and categories 504.

A category 504 is the highest level in the semantic network 222. Broad categories 504 may be created according to a specific industry which fully subsume other knowledge-concepts 502 and skills 404. The semantic network 220 categorizes all knowledge-concepts 502 into categories 504. Knowledge-concepts 502 comprises the next level in the semantic network 220 hierarchy. Each knowledge-concept 502 is a collection of skills 404 that add to the body of knowledge. The semantic network 220 categorizes all skills 404 into knowledge-concepts 502. As described earlier with reference to Figure 4, skills 404 are generic and language independent from all related terms 402. The semantic network 220 categorizes all terms 402 into skills 404. As described earlier with reference to Figure 4, terms 402 comprise

language dependent strings that are found in the electronic document 104. Terms 402 comprise the lowest level in the semantic network 220 hierarchy.

The entire semantic network 220, separate from the thesaurus 221, comprises language independent knowledge that is arranged as a taxonomy. Preferably, the relationships between skills 404 and knowledge-concepts 502 as well as the relationships between knowledge-concepts 502 and categories 504 are many to many. In other words, a single knowledge-concept 502 can comprise several skills 404 and a single skill 404 can be linked to several knowledge-concepts 502. Similarly, several knowledge-concepts 502 may comprises a category 504 and several categories may have links to a single knowledge-concept 502.

To illustrate the terms 402, skills 404, knowledge-concepts 502, and categories 504 of a semantic network 220, the two concepts discussed earlier with reference to Figure 4, namely 'Visual C++' and 'Java', will be used. Both these skills 404 may be grouped under a knowledge-concept 502 'Object Oriented programming languages'. Additionally, the skill 404 'Visual C++' may also belong to the knowledge-concept 502 'Visual Programming Environment'. The knowledge-concept 502 "Visual Programming Environment" may also be linked to other skills 404 such as 'Visual Basic'.

The semantic network 220 uses subsumption as the basis for the hierarchical organization of skills 404, knowledge-concepts 502, and categories 504. In other words, the relationship between skills 404 and knowledge-concepts 502 and knowledge-concepts 502 and categories 504 in the semantic network 220 are based on conceptual subsumption, where a more general object 'subsumes' a more specific object. The concept of subsumption is more general than the concept of synonymy. An object is subsumed by another object if the

subsuming object is much more general than any other subsumed objects and effectively summarizes the subsumed objects. Truly synonymous objects mutually subsume each other. If only synonymous based relationships are allowed, then the granularity between the objects cannot be captured effectively as there are not many truly synonymous objects. The difference between the shades of meaning will not allow correct retrieval in a synonym-based network. The subsumption-based network removes these drawbacks and aids in retrieving related concepts more accurately, since a subsumption is more general compared to a synonym. For example, the object 'JDBC' is subsumed by a more general object called 'Java Programming Language' (a knowledge-concept 502), which is further subsumed by an even more generic object 'Software Engineering' (a category 504).

An object may also be subsumed by more than one higher level object. For example, the skill 404 'JDBC' may be subsumed by at least two knowledge-concepts 502 such as 'Java Programming Language' and 'Database Connectivity Library'. Each of these knowledge-concepts 502 may in turn be subsumed by several categories 504. Hence, the conceptual subsumption also allows many-to-many relationships between skills 404 and knowledge-concepts 502 and between knowledge-concepts 502 and categories 504.

Referring now to Figure 6, a flowchart of the steps of a preferred embodiment of a method performed by the content analysis and semantic network engine 216 is shown. First, identification heuristics are performed (602) on the electronic document 104 to identify the beginning and end of the known sections of interest. The sections of interest are configured by the user when the extraction server 108 is first installed. The sections are then analyzed (604) and information is extracted from the sections. The extracted information is stored (606) in a predefined structure in the target database 110. Using the semantic network 220,

words and word groups are analyzed (608) and the relationships between the different words and word groups are determined and stored in the target database 110. Thus, the present invention advantageously extracts meaningful information from electronic documents, and stores them in a predefined structure in a target database. The extracted information stored in the target database can then be retrieved and manipulated by computer program applications accessing the database. Moreover, the present invention provides a powerful semantic network and thesaurus for defining terms, concepts, meta-concepts, and categories and the relationship between and among such terms, concepts, meta-concepts, and categories. Thus, the semantic network can stored information relating to any field, industry or technology, and allows the extraction server 108 to process various types of documents pertaining to such fields, industries or technologies.

The section processors 218 extract information from sections of interest in an electronic document 104. The particular sections of interest from which information is extracted is determined by the document type. The content analysis and semantic network engine 216 comprises a section processor 218 for extracting words or word groups from each section of interest in an electronic document.

Section processors 218 are configured to operate on a specific document type and may contain one or several section processors 218. For example, resumes typically contain several sections such as a cover letter, contact information, an objective section, an experience section, an education section, a patents section, a publications section, an awards and honors received section, and a courses attended section. In a preferred embodiment, section processors 218 for a resume document type may comprise a cover letter section processor for extracting information from a cover letter, a contact information section processor for

extracting contact information for a candidate, a skills and experience section processor for extracting the skills and experience of a candidate, an education section processor for extracting educational information from a candidate, an awards and honors section processor for extracting any awards and honors received by a candidate, a patents section processor for extracting information about patents obtained by a candidate, and a publications section processor for extracting any articles or documents published by a candidate. Each section processor 218 analyzes a particular section in the electronic document 104 and extracts specific words and word groups from that section into a specific record in the target database 110. Additionally, as described in more detail in commonly assigned U. S. Patent Application Serial No. 09/380,219 entitled "Xtraction Server" by Prabhat K. Andleigh, Nagaraju Pappu, and Vasudeva Kalidindi, each section processor 218 applies a set of heuristics to the particular section of interest in order to analyze and extract the desired information.

Referring now to Figure 7, there is shown a preferred embodiment of the present invention comprising a skills and knowledge information extractor 702. The skills and knowledge information extractor 702 allows the system to automatically extract from a document, such as a resume, the skills of a candidate, the candidate's knowledge in a particular area, and to determine the proficiency level of the candidate in any given skill. Thus, the skills and knowledge information extractor 702 allows a user to automatically determine a "career profile" of a candidate from his or her resume. As used herein, a "career profile" refers to any qualitative and quantitative information about a candidate's work history, experience, and proficiency. For example, such information includes, but is not limited to, how long a candidate worked in a particular profession, when, where, and at what

depth did the candidate gain experience in a particular skill, what is the candidate's overall knowledge level in a particular area, how much management experience a candidate has, etc.

As used herein, "terms" refers to the actual word or words which are found in a resume, "skill" or "skill information" refers to the skills 404 in the thesaurus 221 and semantic network 220 which relate to those terms, and "knowledge" or "knowledge information" refers to the knowledge-concepts 502 relating to the skills. For example, in a resume, a candidate may have used the terms "Microsoft Visual C++" or "MS VC++". The present invention would identify these terms as belonging to the skill "C++", which in turn is related to the knowledge "object oriented programming" which in turn may be related to the category "Software." Thus, although the only terms actually used in and extracted from the resume were "Microsoft Visual C++" and "MS VC++", the present invention is able to determine that the candidate has "skill" in C++ and has "knowledge" of object oriented programming even though the words C++ and object oriented programming were never used in the document.

The skill and knowledge information extractor 702 uses a non-monotonic reasoning principle to determining a candidate's skill level. As described above, non-monotonic reasoning refers to the use of default assumptions which are made about the state of unknown factors. These default assumptions may be changed as new information or evidence becomes available. Additionally, default assumptions may be changed due to the absence of certain information or evidence. The operation of the non-monotonic reasoning approach used by the skill and knowledge for information extractor 702 is best illustrated using an example.

During operation, the present invention finds a skill, X, in a candidate's resume, R. In the absence of any other knowledge, the skill and knowledge information extractor 702

assumes that the skill level of the candidate for the skill X is average. As the skill and knowledge information extractor 702 obtains additional information from the resume R about skill X, the assumption of the skill level for skill X is refined. Additional knowledge that may be used to refine the skill level includes, but is not limited to, the section in which the skill X is found. For example, if the skill X is found in the Objective Section of a resume R, a positive numerical value, or objective weightage factor W(O), will be added to the skill level. Additionally, a positive weight for each project in which the skill X is used, represented here by $W(P_i)$, may be added to the skill level. Preferably, this weightage value is computed for all projects in resume R. The number of associated skills that are also used, W(K), may also be added to the skill level. As used herein, associated skills are the skills related to the main skill; knowing a main skill implies that a person also knows all associated skills. For example, if one is an expert in the skill "database programming" or "database administration," this person must be knowledgeable in the associated skill "SQL." Associated skills can be determined using the semantic network 220 and the thesaurus 221. For a given skill x, all its associated skills ($X_1 \ldots X_n$) are linked with x through the semantic network 220 and thesaurus 221. For example, a thesaurus 221 entry for the "skill database administration" would contain links to the "skills database server administration," "database user management," and/or "SQL." Also, the number of years of experience for the skill X, W(Y) may also be added to the skill level. Moreover, the number of years since the skill X was used may represent a negative factor, W(LU), which is subtracted from the skill level. Thus, in a preferred embodiment, a summation of the weights described above gives a specific skill level for the skill X. A mathematical representation for determining the skill level of a particular skill is as follows:

$$SkillLevel(X') = SkillLevel(X) + W(O) + W(P_i) + W(K) + W(Y) - W(LU)$$

The weightage functions are computed using the total number of skill levels that are defined, and the distance from the current skill level to the next skill level. One skilled in the art will realize that the weightage factors used to adjust the skill level are not limited to those listed in the above example but can comprise any number of factors to be determined by the system creator.

The computation of the skill level of a particular skill for a candidate can also be demonstrated using an example. Initially, the skill and knowledge information extractor 702 assumes that a person has an average skill level for a particular skill such as C++. If the candidate's resume states that the candidate took a course in C++, that fact would add a positive weightage factor to the skill level, thus adjusting the average skill level to a higher value. If the candidate's resume also states that the candidate has two years of work experience in C++, that fact would add another positive weightage factor to the skill level and adjust the average skill level to another higher value. The values by which the average skill level is adjusted for the C++ course and the two years of work experience are not necessarily the same but may reflect the value attributed by the system creator. Each mention of C++ in the resume would this be used to adjust the skill level either up or down. Additionally, the user of terms in the resume which are related in the semantic network 220 and thesaurus 221 to the concept or skill C++ could also be used to adjust the skill level of the candidate. After all the relevant terms in the candidate's resume have been extracted and evaluated, the skill

and knowledge information extractor 702 determines a single value for the skill level for the candidate for the particular skill.

After a final skill value for a particular skill has been determined, the skill and knowledge information extractor 702 then maps the skill value to a scale for qualitatively illustrating the proficiency of the candidate in that particular skill. For example, if a final skill value for a particular candidate has been determined to be the number 6.8, that number may map to a rating of "good" on a scale of 1 to 10, with 1 being poor and 10 being excellent. Thus, the present invention allows a user to determine the proficiency of a candidate's skill level for a particular skill and to ascribe a qualitative value to that proficiency level. One skilled in the art will realize that the qualitative scales used to describe a particular skill value may be any type of scale with a range of numerical values and/or adjective descriptors. For example, a qualitative scale may map the final skill value to a scale comprising numbers such as 1 to 5 or 1 to 10. A scale may map the final skill value to a scale comprising numbers and adjectives such as 1 (poor) to 10 (excellent). The qualitative scale may be determined by the system creator.

The categories, knowledge, skills and terms are preferably set up in a relational database prior to the extraction process. As described above with reference to Figures 4 and 5, in a preferred embodiment, the relationship between categories and knowledge is many-many, the relationship between knowledge and skills is many-to-many, and the relationship between skills and terms is one-to-many.

Referring now to Figure 8, there is shown a flow chart of a preferred embodiment of a method for the present invention. In a preferred embodiment, a resume is evaluated (802) for a particular skill. The skill level for that particular skill is then determined (804) using the

above described techniques. After a final skill level value is determined, the skill level is mapped (806) to a qualitative scale. Finally, the skill and the qualitative scale value of the skill level is stored (808) in the target database. More specifically, the categories, knowledge, skills and terms (i.e. the semantic network) are loaded into main memory. The electronic document text is then passed to the skill and knowledge information extractor 702. In a preferred embodiment, knowledge, skills, skill levels and number of years are extracted from the electronic document in the following manner: first, all the terms in the database are checked against the document, then an initial scan of the document collects all the terms. The frequency of appearance of the term is recorded. Afterwards, the weightage factors for the skill level calculation are applied. A second scan of the electronic document analyses the document and a running list is maintained for all terms to calculate the experience duration where the term is maintained. On completion of the second scan, all the terms are rolled up into skills according to the semantic network and thesaurus, all the skills are rolled up into knowledge according to the semantic network, and all the knowledge items are rolled up into categories. Additionally, categories specifically mentioned are added. Thus, based on this information, the skill levels and years of experience are computed as described above.

Referring now to Figure 9, there is shown a screen shot of a user interface of a preferred embodiment of a target database for a skill and knowledge information extractor 702. Window 902 displays the particular skills analyzed from a candidate's resume, the qualitative level determined by the skill and knowledge information extractor 702, and the years of experience the candidate has for the particular skill. For example, the highlighted portion of window 902 indicates that the candidate has some skill as an analyst, that the qualitative proficiency of the candidate's skill as an analyst is "excellent", and that the

candidate has 4 years of experience as an analyst. Thus, the present invention advantageously allows a user to extract, determine, and display from a candidate's resume the proficiency of a particular skill of the candidate.

The present invention is designed as a set of Object Oriented Libraries and contains the following major Object Libraries:

Xpert Object Library	This library encapsulates the Xerox InXight APIs and provides basic building blocks of extractions. For example, the date Xpert Object can decide whether a given text contains a date or not.
DataBase Library	This library encapsulates the ODBC APIs to connect to the target database, as well as provides Objects for creation and manipulation of extracted records.
Document Filter Library	This library provides Objects that can filter the input document, decide the document type and formatting. For example, the Word Document Object can decide if the input document is a Microsoft Word document or not and can extract the text from the Word files.
Paragraph Property Library	This library provides Objects that provide mechanisms to gather the paragraphs from the input document, Heuristics Objects which gather the Paragraph Properties from the input document, etc.... This information is later used in the extraction process.
Extraction Object Library	This library mainly contains one Object for each of the sections typically present in a document. For example, for a Resume document, it has a Resume Object, Section Object, Experience, Education, HonorsandAwards, Publications, Patents, Objective, References, CoverLetter, and Contact Information Objects. Each of these Objects has all the logic necessary to extract content from that particular section in a Resume.
Knowledge and Skill Level Object Library	This Library provides objects and facilities for extractive Knowledge of a person from a resume and provides objects and functions to calculate the skill level of a particular skill using the procedure described in the previous sections.
Thread Library	This library provides Objects for multi-threading and preferably encapsulates the Win32 thread API. It provides mechanisms for thread synchronization, semaphores, lock and resource management.

In a preferred embodiment, the present invention may be implemented to run on a Windows NT Server and any relational database such as Oracle Database. Database tables may be used to define how information is represented in a relational or object-oriented database. In an object-oriented implementation, any relational table is preferably represented as an object class. The following section describes a preferred embodiment of the content and type of the fields that are extracted into a relational database, and also the definitions of the categories, knowledge, skills and terms. The supporting tables are also explained. One skilled in the art will realize that these tables are not limited to the specific information illustrated therein but may be created as needed, depending on the document type being processed.

Table 1

AutoEntryDocuments

Table 1 holds the documents that are to be extracted. It holds the following information:

DocID	Document ID
AEDocFileName	The complete path and file name of the document
AerfCategory	Category of the Resume/Document
Task_ID	TaskID associated with the document
AEStatus	Status of Extraction. (Not Done, Done, Errors)

Table 2

AutoEntrySchedule

Table 2 holds information about the scheduled extraction tasks.

Task_ID	Task ID of the Schedule
AEScheduleDate	Date the Task is Scheduled to Run
AESrfCategory	Category of the documents scheduled for this Task
AEActualDate	Date the Task is scheduled to Run
. . AEScheduleType	Type of the Schedule (Daily, Weekly, Monthly)
AEScheduleStatus	Status of the Task (Scheduled, Completed, Errors)

Table 3

Candidate

Table 3 holds the personal information like name of the person, contact address, current employer, resume summary etc. The XtractionXpert automatically extracts the following information from the resume:

CAFirstName	First Name of the Person
CALastName	Last Name of the Person
CANickName	Nick Name of the Person
CAWorkCompany	Current Employer
CATitle	Current Designation
CAYearsEmployed	Total Number of years of experience
CandidateID	Candidate ID, the database ID of the person
CASalutation	Salutation (Mr, Ms, Dr etc.)
CACurrentObjective	Stated Objective in the Resume
CABriefExperience	Text of the Summary Section
CAYearsOfExperienc	Years of Experience with current employer
CALastModifiedDate	Date of Last Modification to the Record
CATextResume	Actual text of the resume
CAModifiedBy	Person who modified the record
CAWorkMailStop	Mail Stop of the Work Address
CAWorkPhoneNo	Work Phone Number
CAWorkExtension	Work Phone Number Extension
CAWorkFaxNo	Work Fax Number
CAWorkMobilePhone	Work Related Mobile Phone Number
CAWorkEmail	Official Email Address
CAHomeMailStop	Mail Stop of Residence
CAHomePhoneNo	Home Phone Number
CAHomeExtension	Home Phone Extension
CAHomeFaxNo	Home Fax Number
CAHomeEmail	Home Email Address
CAOtherMailStop	Mail Stop other than Office and Residence
CAOtherPhoneNo	Any other Phone Number (e.g. Recruiting Agency)
CAOtherExtension	Extension number of Phone
CAOtherFaxNo	Fax Number other than work and residence
CAOtherMobilePhone	Mobile Phone Number
CAOtherEmail	Email Address other than residence and work
CAWorkStreet	Street Name of the Work address
CAWorkSuiteNo	Suite Number of Work address
CAWorkCity	City Name of the Work Address
CAWorkState	State Name of the Work Address
CAWorkZip	Zip Code of the Work Address
CAWorkCountry	Country of the Work Address
CAHomeStreet	Street Name of the Home Address
CAHomeSuiteNo	Suite/Apt. Number of the Home
CAHomeCity	City of Residence Address
CAHomeState	State of Residence Address
CAHomeZip	Zip Code of the Residence Address
CAHomeCountry	Country of the Residence Address
CAOtherStreet	Street Name of address other than work and home
CAOtherSuiteNo	Suite or Apt. Number other than work and phone

CAOtherCity	Name of City from the other address
CAOtherState	State Name of the other address
CAOtherZip	Zip Code of the other address
CAOtherCountry	Country of the other address
CAHomeMobilePhone	Mobile phone of other address
CAHomePage	Web address
CAPager	Pager Number

Table 4

ExperienceDetail

CandidateID	Database ID of the Person (Candidate Table)
EDEmployerName	Name of the Company worked for
EDReportedTo	Name of the reporting Manager
EDResponsibilityL1	Primary Responsibility (Designation)
EDResponsibilityL2	Secondary Responsibility
EDPeopleManaged	Number of people managed
. . EDHighlightsl	First Bulleted item from the Experience Description
EDHighlights2	Second Bulleted item
EDHighlights3	Third Bulleted Item
EDHighlights4	Fourth Bulleted Item
EDNotes Text of the	Experience Description
ExperienceDetailID	ID ofthe Record
EDStartDateDD	Date Joined for the company
EDStartDateMM	Month joined for the company
EDStartDateYYYY	Year joined for the company
EDEndDateDD	Date last worked for the company
EDEndDateMM	Month last worked for the company
EDEndDateYYYY	Year last worked for the company
EDReportedToPhone	Manager's Phone Number

5

Table 5

EducationRecord

CandidateID	Database ID of the person
EdrfMajor	Specialization
EDAwardedby	Name of the Institution
EDGPA	GPA earned
EDNote	Text of the description of the record
EdrfGradStatus	Status of graduation (passed, pending etc.)
EdrfDegreeType	Type of the Degree (B.S., M.S, Ph.D)
EDStartDateDD	Date joined in the course
EDStartDateMM	Month of joining
EDStartDateYYYY	Year of joining
EDEndDateDD	Date of Completion

EDEndDateMM	Month of Completion
EDEndDateYYYY	Year of Completion

Table 6
PatentRecord

CandidateID	Database Id of the Person
PATitle	Title of the Patent
Pacountry	Country where Patent was filed
PAJointHolder	Name of the Joint Holder
PAPatentNumber	Patent Number
PAGrantDateYYYY	Year Patent Granted
PAPatentStatus	Status of the Patent (Granted, Pending)
PANotes	Text of the description of the Patent
PAGrantDateMM	Month Patent granted
PAGrantDateDD	Date Patent granted

Table 7
PublicationRecord

CandidateID	Database Id of the Person
PurfPublicatType	Type of Publication (Book, Paper etc.)
PUTitle	Title of Publication
PUPublicationName	Name of the Publication
PUDateDD	Date of Publication
PUPublisherName	Name of the Publisher
PUDateMM	Month of Publication
PUPageRange	Page Numbers
PUDateYYYY	Year of Publication
Puisbn	ISBN number of the Publication
PUNotes	Text of the description

Table 8
SkillRecord

CandidateID	Database Id of the Person
ConceptID	Pointer to Concept Table where Skill Name is found
SKEXpYears	Number of years of experience in the skill
SKNotes	description of the skill
SkrfProfLevel	Skill Level

Table 9

Kno wledgeRecord

CandidateID	Database ID of the Person
KNYear	Number of years of experience
MetaConceptID	Pointer to MetaConcept
KNComment	Any comments associated

Table 10

ProfAssocRecord

CandidateID	Database ID of the person
PRAssocName	Name of the Professional Body
PRMemberCategory	Type of the membership
PRAssociatedSinceDD	Date of joining
PRAssociatedSinceMM	Month of joining
PRAssociatedSinceYYYY	Year of joining

Table 11

ProfLicenseRecord

CandidateID	Database ID of the person
PLLicenceName	Name of the License
PLLicenceAuthority	Name of the organization that issued the license
PLLicenceNumber	Professional License Number
PLLicenceGranted	Whether License Granted
PLLicenceLevel	Level of the License
PLLicenceState	Current status of the license
PLNotes	Text of the description
PLExpirationDateDD	Expiration Date (Date)
PLExpirationDateMM	Expiration Date (Month)
PLExpirationDateYYYY	Expiration Date (Year)

Table 12

ReferenceRecord

RECandidateRole	Role Played by the Candidate in the Team
CandidateID	Database ID of the Candidate
RERefrenceName	Name of the Referee
REReferenceTitle	Title (Designation) of the Referee
REWorkPhoneNo	Work Phone Number of the Referee
REHomePhoneNo	Home Phone Number of the Referee
RECandidateRelation	Relationship of the Referee to the Candidate
RECandidateDateBegin	Date candidate started the associationship
RECandidateDate End	Date association ended

Table 13

Courses

COCourseName	Name of the Course taken
CandidateID	Database Id of the Candidate
CODDateDD	Date course taken (date)
CODateMM	Date course taken (month)
CODateYYYY	Date course taken (year)
CoNotes	Description of the course

Table 14

AwardsHonors

Awhighlight	Name and highlight of the Award or Honor
CandidateID	Database Id of the candidate
AWNotes	Description of the Award or Honor

Table 1

MiscellenousInformation

MINotes	Text of the any other section
CandidateID	Database Id of the candidate

Table 16

Category

Table 16 provides information regarding the relationships between categories and knowledge information.

MetaConceptID	Knowledge ID
CmrfCategory	Category ID

Table 17

MetaConcept

Table 17 provides knowledge information for semantic network 220.

MetaConceptID	Knowledge ID
MEMetaConceptName	Name of the Knowledge Entry
Medescription	Description of the Knowledge Entry
MESemMarer	Semantic Markers and Types

Table 18

Concept

Table 18 provides information relating to skills.

ConceptID	Skill ID
CNConceptName	Name of the Skill
CNDescription	Description of the Skill
CNSemMarer	Semantic Markers and Types

Table 19

ConceptRelation

Table 19 provides information on relationships between skills and knowledge.

MetaConceptID	Knowledge ID
ConceptID	Skill ID
CRRelationType	Type of the Relation between Knowledge and Skill
CrisaRelationYN	Specifies if the relation is hierarchial
CRDescription	Description of the relation

Table 20

Term

Table 20 provides information on terms.

TermID	Term ID
TETerm	Name of the Term
LanguageID	Language ID of the Term
ConceptID	Skill ID to which the Term belongs

Table 21

Language

Table 21 stores information about different languages to which the terms belong.

LanguageID	Language ID
LALanguage	Name of the Language (English, French etc.)

Table 22

CaWord

CWWordID	Word ID
CWClassification	Classification of the word
CWWord	Word found

Table 23

CaWordList

CWWordID	Word ID

CWLFirstDocNumber	The First Document in which the word was found
CWLBlock	Block which consists the document Ids
CWLFlag	Database Flag

Table 24
CaWordPosition

CWWordID	Word ID
CWPFistDocNumber	First Document in which the word was found
CWPFlag	Database Flag
CWPBlock	Block which consists of document ids

From the above description, it will be apparent that the invention disclosed herein provides a novel and advantageous system and method for extracting and analyzing skill and knowledge information from an electronic document. The foregoing discussion discloses and describes merely exemplary methods and embodiments of the present invention. As will be understood by those familiar with the art, the invention may be embodied in other specific forms without departing from the spirit of the invention or essential characteristics thereof. Accordingly, the disclosure of the present invention is intended to be illustrative, but not limiting, of the scope of the invention, which is set forth in the following claims.

6. The apparatus of claim 2 wherein a skill extracted from the electronic document and the skill mapping to a qualitative scale are displayed on a computer.

7. An apparatus for analyzing and extracting skill and knowledge information from an electronic document into a target database having predefined fields, the apparatus comprising:

a thesaurus for linking together terms and skills and for defining relationships between and among the terms and skills; and

a semantic network coupled to the thesaurus for organizing terms and skills in the thesaurus, knowledge, and categories in a hierarchical structure;

wherein the thesaurus and semantic network are used to analyze skill and knowledge information in the electronic document.

8. The apparatus of claim 7 further comprising:

a document pre-processor coupled to the semantic network for classifying the electronic document as a document type and for performing an initial analysis on the electronic document.

9. The apparatus of claim 7 further comprising:

a heuristics engine coupled to the semantic network for applying a set of heuristics to the electronic document.

10. The apparatus of claim 7 further comprising:

mapping the skill level to a qualitative scale.

16. A computer implemented method for extracting and displaying skill and knowledge information from an electronic document, the method comprising the steps of:

identifying skill and knowledge information in the electronic document;

extracting the skill and knowledge information from the electronic document;

determining a skill level for skill information extracted from the electronic document; and

mapping the skill level to a qualitative scale.

17. A computer-readable medium for extracting and displaying skill and knowledge information from an electronic document, the computer-readable medium comprising code for performing the steps of:

identifying skill and knowledge information in the electronic document;

extracting the skill and knowledge information from the electronic document;

determining a skill level for skill information extracted from the electronic document; and

mapping the skill level to a qualitative scale.

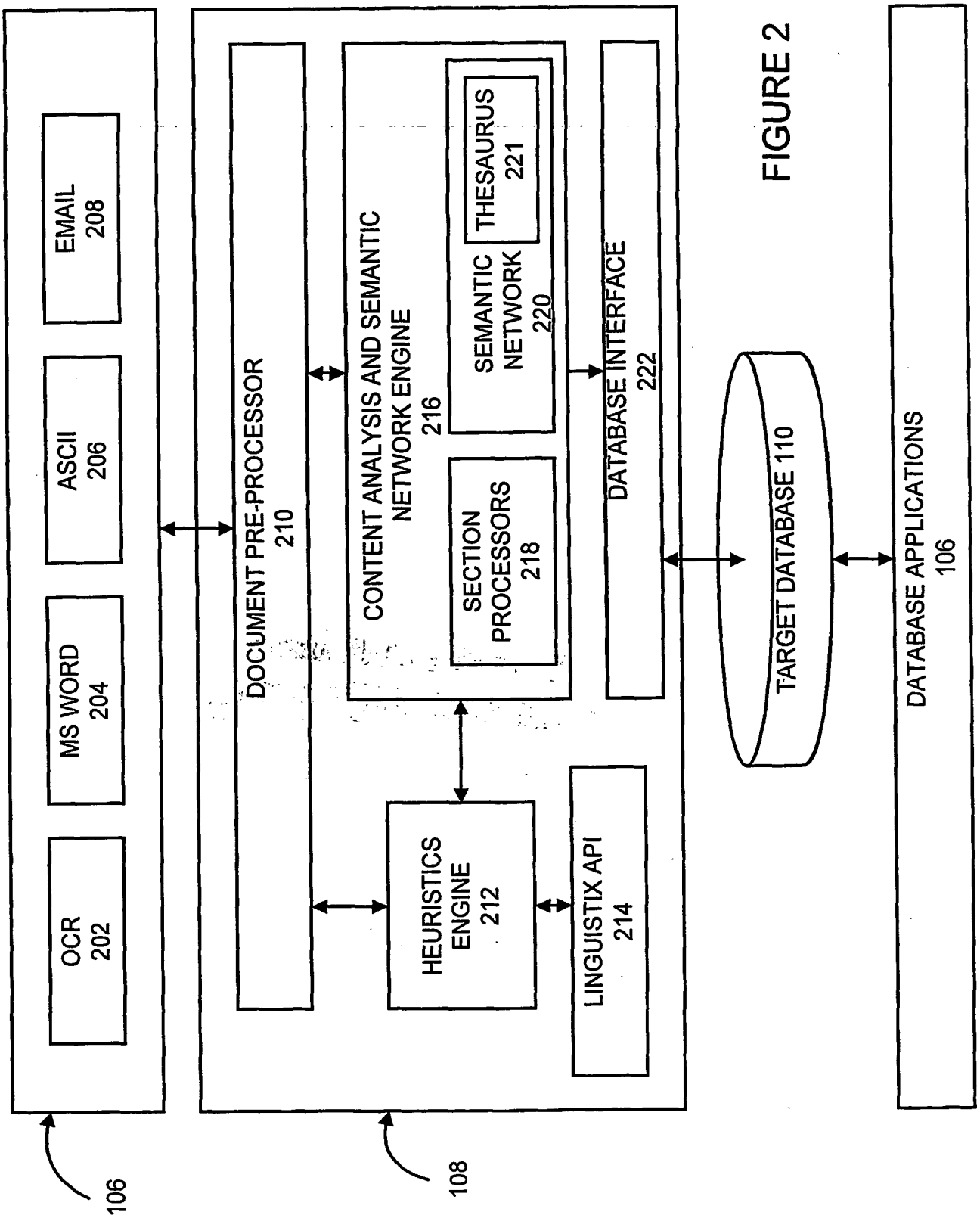


FIGURE 2

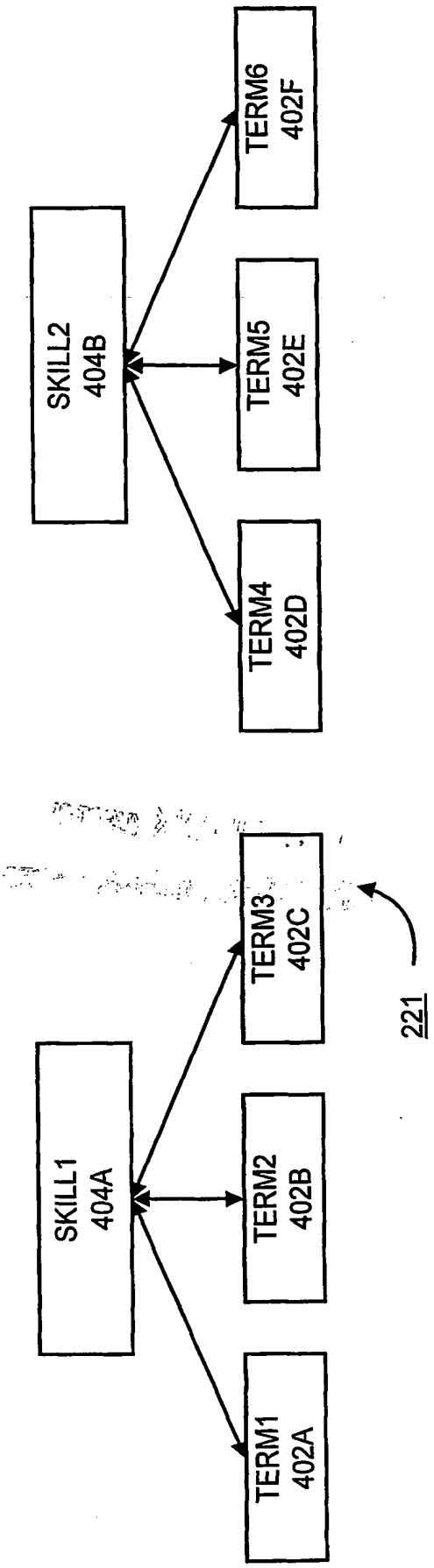


FIGURE 4

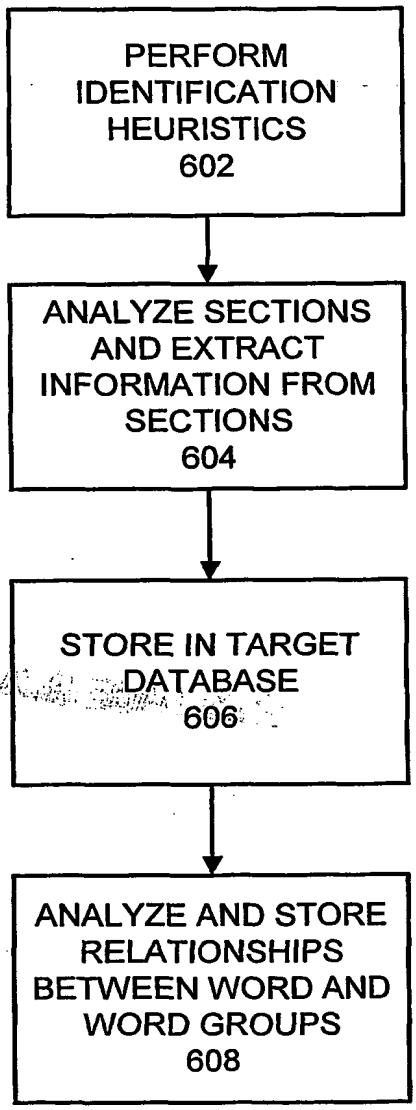


FIGURE 6

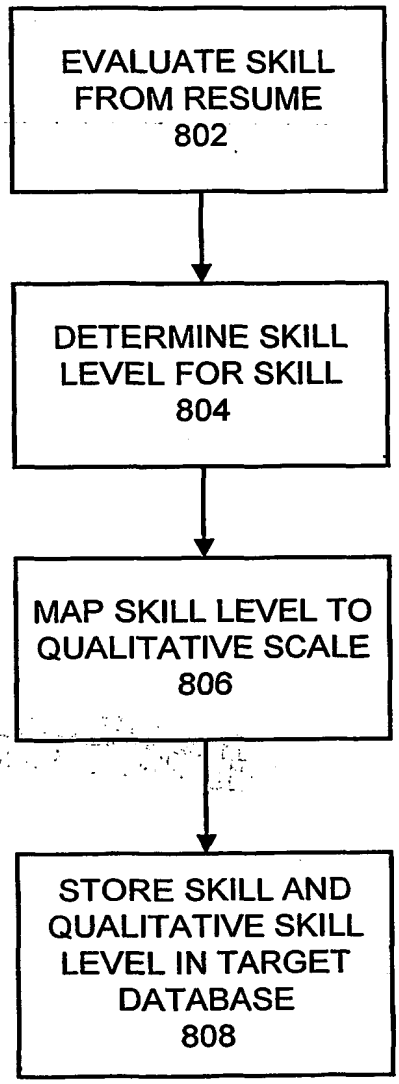


FIGURE 8

INTERNATIONAL SEARCH REPORT

ıational Application No
PCT/US 99/26083

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/60

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 197 004 A (SOBOTKA DAVID ET AL) 23 March 1993 (1993-03-23) abstract column 3, line 38 -column 3, line 59 column 4, line 22 -column 4, line 68 column 5, line 49 -column 6, line 40 column 7, line 10 -column 8, line 58 claims figures 1,2,5-8 ---	1-17
Y	US 5 297 039 A (KANAEGAMI ATSUSHI ET AL) 22 March 1994 (1994-03-22) abstract column 2, line 7 -column 6, line 59 figures 3,5,6 --- -/--	1-17

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
31 March 2000	07/04/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Abbing, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

ıational Application No
PCT/US 99/26083

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5197004 A	23-03-1993	NONE	
US 5297039 A	22-03-1994	JP 2943447 B JP 4357568 A	30-08-1999 10-12-1992
WO 9839716 A	11-09-1998	AU 6182098 A	22-09-1998
US 5416694 A	16-05-1995	CA 2140216 A	29-08-1995